

Bayesian Manifold Regression

Yun Yang ^{*} and David B. Dunson

*Department of Statistical Science
Duke University
Box 90251
NC 27708-0251, Durham, USA
e-mail: yy84@stat.duke.edu*

dunson@stat.duke.edu

Abstract: There is increasing interest in the problem of nonparametric regression with high-dimensional predictors. When the number of predictors D is large, one encounters a daunting problem in attempting to estimate a D -dimensional surface based on limited data. Fortunately, in many applications, the support of the data is concentrated on a d -dimensional subspace with $d \ll D$. Manifold learning attempts to estimate this subspace. Our focus is on developing computationally tractable and theoretically supported Bayesian nonparametric regression methods in this context. When the subspace corresponds to a locally-Euclidean Riemannian manifold, we show that a Gaussian process regression approach can be applied that leads to the minimax optimal adaptive rate in estimating the regression function under some conditions. The proposed model bypasses the need to estimate the manifold, and can be implemented using standard algorithms for posterior computation in Gaussian processes. Finite sample performance is illustrated in an example data analysis.

AMS 2000 subject classifications: Primary 62H30, 62-07; secondary 65U05, 68T05.

Keywords and phrases: Asymptotics, Contraction rates, Dimensionality reduction, Gaussian process, Manifold learning, Nonparametric Bayes, Subspace learning.

1. Introduction

Dimensionality reduction in nonparametric regression is of increasing interest given the routine collection of high-dimensional predictors in many application areas. In particular, our primary focus is on the regression model

$$Y_i = f(X_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n, \quad (1.1)$$

where $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^D$, f is an unknown regression function, and ϵ_i is a residual having variance σ^2 . We face problems in estimating f accurately due to the moderate to large number of predictors D . Fortunately, in many applications, the predictors have support that is concentrated near a d -dimensional subspace \mathcal{M} . If one can learn the mapping from the ambient space to this subspace, the

^{*}Supported by grant ES017436 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH).

dimensionality of the regression function can be reduced massively from D to d , so that f can be much more accurately estimated.

There is an increasingly vast literature on the topic of subspace learning, but there remains a lack of approaches that allow flexible non-linear dimensionality reduction, are scalable computationally to moderate to large D , have theoretical guarantees and provide a realistic characterization of uncertainty. Regarding this last point, we would like to be able to characterize uncertainty in estimating the regression function f , in functionals of f of interest and in predictions. Typical two-stage approaches in which one conducts dimensionality reduction in a first stage, and then plugs the d -dimensional features into a next stage regression may provide a point estimate with good properties but do not characterize uncertainty in this estimate.

With this motivation, we focus on Bayesian nonparametric regression methods that allow \mathcal{M} to be an unknown Riemannian manifold. One natural direction is to choose a prior to allow uncertainty in \mathcal{M} , while also placing priors on the mapping from x_i to \mathcal{M} , the regression function relating the lower-dimensional features to the response, and the residual variance. Some related attempts have been made in the literature. [29] propose a logistic Gaussian process model, which allows the conditional response density $f(y|x)$ to be unknown and changing flexibly with x , while reducing dimension through projection to a linear subspace. Their approach is elegant and theoretically grounded, but does not scale efficiently as D increases and is limited by the linear subspace assumption. Also making the linear subspace assumption, [23] proposed a Bayesian finite mixture model for sufficient dimension reduction. [22] instead propose a method for Bayesian nonparametric learning of an affine subspace motivated by classification problems.

There is also a limited literature on Bayesian nonlinear dimensionality reduction. Gaussian process latent variable models (GP-LVMs) [16] were introduced as a nonlinear alternative to PCA for visualization of high-dimensional data. [15] proposed a related approach that defines separate Gaussian process regression models for the response and each predictor, with these models incorporating shared latent variables to induce dependence. The latent variables can be viewed as coordinates on a lower dimensional manifold, but daunting problems arise in attempting to learn the number of latent variables, the distribution of the latent variables, and the individual mapping functions while maintaining identifiability restrictions. [8] instead approximate the manifold through patching together hyperplanes. Such mixtures of linear subspace-based methods may require a large number of subspaces to obtain an accurate approximation even when d is small.

It is clear that probabilistic models for learning the manifold face daunting statistical and computational hurdles. In this article, we take a very different approach in attempting to define a simple and computationally tractable model, which bypasses the need to estimate \mathcal{M} but can exploit the lower-dimensional manifold structure when it exists. In particular, our goal is to define an approach that obtains a minimax-optimal adaptive rate in estimating f , with the rate adaptive to the manifold and smoothness of the regression function. Surprisingly,

we show that this can be achieved with a simple Gaussian process prior.

Section 2 defines the proposed model and gives some basic geometric background along with a heuristic motivation for the model. Section 3 contains a more thorough background of necessary geometric concepts. Section 4 provides theory on adaptive rates. Section 5 contains simulation studies of finite sample performance relative to competitors, and Section 7 discusses the results.

2. Gaussian Processes on Manifolds

2.1. Background

Gaussian processes (GP) are widely used as prior distributions for unknown functions due to tractable posterior computation and strong theoretical guarantees. For example, in the nonparametric regression (1.1), a GP can be specified as a prior for the unknown function f . In classification, the conditional distribution of the binary response Y_i is related to the predictor X_i through a known link function h and a regression function f as $Y_i|X_i \sim \text{Ber}[h\{f(X_i)\}]$, where f is again given a GP prior. The following developments will mainly focus on the regression case. The GP with squared exponential covariance is a commonly used prior in the literature. The law of the centered squared exponential GP $\{W_x : x \in \mathcal{X}\}$ is entirely determined by its covariance function,

$$K^a(x, y) = EW_x W_y = \exp(-a^2 \|x - y\|^2), \quad (2.1)$$

where the predictor domain \mathcal{X} is a subset of \mathbb{R}^D , $\|\cdot\|$ is the usual Euclidean norm and a is a length scale parameter. Although we focus on the squared exponential case, our results can be extended to a broader class of covariance functions with exponentially decaying spectral density, including standard choices such as Matérn, with some elaboration. We use $GP(m, K)$ to denote a GP with mean $m : \mathcal{X} \rightarrow \mathbb{R}$ and covariance $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$.

Given n independent observations, the minimax rate of estimating a D -variate function that is only known to be Hölder s -smooth is $n^{-s/(2s+D)}$ [26]. A function in \mathbb{R}^D is said to be Hölder s -smooth if it has bounded mixed partial derivatives up to order $\lfloor s \rfloor$ for $\lfloor s \rfloor$ the largest integer strictly smaller than s with the partial derivative of order $\lfloor s \rfloor$ being Lipschitz-continuous of order $s - \lfloor s \rfloor$. Surprisingly, [31] proved that, for Hölder s -smooth functions, a prior specified as

$$W^A|A \sim GP(0, K^A), \quad A^D \sim Ga(a_0, b_0), \quad (2.2)$$

for $Ga(a_0, b_0)$ the Gamma distribution with pdf $p(t) \propto t^{a_0-1}e^{-b_0 t}$ leads to the minimax rate $n^{-s/(2s+D)}$ up to a logarithmic factor $(\log n)^\beta$ with $\beta \sim D$ adaptively over all $s > 0$ without knowing s in advance. The superscript in W^A indicates the dependence on A , which can be viewed as a scaling or inverse bandwidth parameter. Although the sample paths from this GP prior are almost surely infinitely differentiable, an intuitive explanation for such smoothness

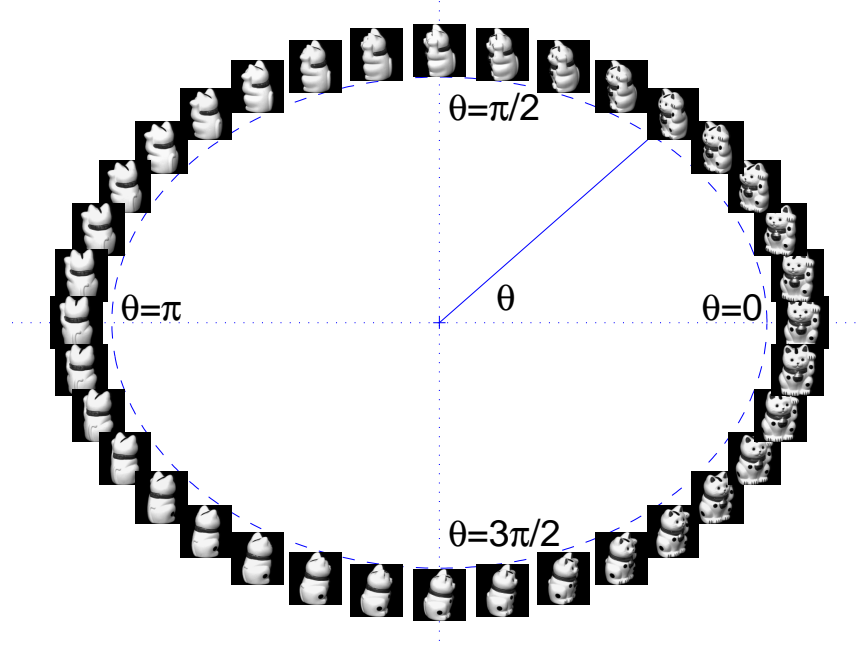


FIG 1. In this data, 72 size 128×128 images were taken for a “lucky cat” from different angles: one at every 5 degrees of rotation. 36 images are displayed in this figure.

adaptability is that less regular or wiggly functions can be well approximated by shrinking the long path of a smooth function by a large factor a .

In many real problems, the predictor X can be represented as a vector in high dimensional Euclidean space \mathbb{R}^D , where D is called the ambient dimensionality. Due to the curse of dimensionality, the minimax rate $n^{-s/(2s+D)}$ will deteriorate rapidly as D increases. This will become extremely fatal in the notorious small n large p problem, where D can be much larger than the sample size n . In such high dimensional situations, there is no hope to accurately estimate the regression function f without any assumption on the true model. One common assumption requires that f only depends on a small number $d \ll n$ of components of the vector X that are identified as important. In the GP prior framework, [25] proposed to use “spike and slab” type point mass mixture priors for different scaling parameters for each component of X to do Bayesian variable selection. [4] showed that carefully calibrated implementations of this approach can lead to minimax adaptive rates of posterior concentration. However, variable selection is a very restrictive notion of dimension reduction. Our focus is on a different notion, which is that the predictor lies on a manifold \mathcal{M} of intrinsic dimension d much lower than the ambient space dimension D . This manifold can be considered as a d dimensional hyper surface in \mathbb{R}^D . A rigorous definition is described in section 3. A concrete example is shown in Fig.1. These data ([21]) consist of 72 images of a “lucky cat” taken from different angles $5^\circ, 10^\circ, \dots$. The predictor

$X \in \mathbb{R}^{128^2}$ is obtained by vectorizing the 128×128 image. The response Y is a continuous function f of the rotation angle $\theta \in [0, 2\pi]$ satisfying $f(0) = f(2\pi)$, such as \sin or \cos functions. Intuitively, the predictor X concentrates on a circle in $D = 128^2$ -dim ambient space and thus the intrinsic dimension d of X is equal to one, the dimension of the rotation angle θ .

2.2. Our Model and Rate Adaptivity

When $X \in \mathcal{M}$ with \mathcal{M} d -dimensional, a natural question is whether we can achieve the intrinsic rate $n^{-s/(2s+d)}$ for f Hölder s -smooth without estimating \mathcal{M} . Surprisingly, the answer is affirmative. [33] showed that a least squares regularized algorithm with an appropriate d dependent regularization parameter can ensure a convergence rate at least $n^{-s/(8s+4d)}(\log n)^{2s/(8s+4d)}$ for functions with Hölder smoothness $s \leq 1$. [5] proved that local polynomial regression with bandwidth dependent on d can attain the minimax rate $n^{-s/(2s+d)}$ for functions with Hölder smoothness $s \leq 2$. However, similar adaptive properties have not been established for a Bayesian procedure. In this paper, we will prove that a GP prior on the regression function with a proper prior for the scaling parameter can lead to the minimax rate for functions with Hölder smoothness $s \leq \{2, \gamma - 1\}$, where γ is the smoothness of the manifold \mathcal{M} . In the remainder of this section, we first propose the model, and then provide a heuristic argument explaining the possibility of manifold adaptivity. Formal definitions and descriptions of important geometric concepts can be found in the next section.

Analogous to (2.2), we propose the prior for the regression function f as

$$W^A | A \sim GP(0, K^A), \quad A^d \sim Ga(a_0, b_0), \quad (2.3)$$

where d is the intrinsic dimension of the manifold \mathcal{M} and K^a is defined as in (2.1) with $\|\cdot\|$ the Euclidean norm of the ambient space \mathbb{R}^D . Although the GP in (2.3) is specified through embedding in the \mathbb{R}^D ambient space, we essentially obtain a GP on \mathcal{M} if we view the covariance function K^a as a bivariate function defined on $\mathcal{M} \times \mathcal{M}$. Moreover, this prior has two major differences with usual GPs or GP with Bayesian variable selection:

1. Unlike GP with Bayesian variable selection, all predictors are used in the calculation of the covariance function K^a ;
2. The dimension D in the prior for inverse bandwidth A is replaced with the intrinsic dimension d .

Generally, the intrinsic dimension d is unknown and needs to be estimated. Many estimation methods has been proposed [7, 6, 17, 18]. For example, [17] considered a likelihood based approach and [18] relies on singular value decomposition of local sample covariance matrix. We will use [17] to obtain an estimator \hat{d} and then plug in this estimator into our prior (2.3) to obtain an empirical Bayes approach.

In our model, we only need to estimate the intrinsic dimensionality d rather than the manifold \mathcal{M} . Most algorithms for learning \mathcal{M} become computationally

demanding as the ambient space dimensionality D increases, while estimating d is fast even when D is tens of thousands. Moreover, although we use the full data in the calculation of the covariance function, computation is still fast for moderate sample sizes n regardless of the size of D since only pairwise Euclidean distances among D -dimensional predictors are involved whose computational complexity scales linearly in D . This dimensionality scalability provides huge gains over two stage approaches (section 2.3) in high dimensional regression settings even though they can also achieve the optimal posterior convergence rate (Theorem 2.3).

Intuitively, one would expect that geodesic distance should be used in the square exponential covariance function (2.1). However, there are two main advantages of using Euclidean distance instead of geodesic distance. First, when geodesic distance is used, the covariance function may fail to be positive definite. In contrast, with Euclidean distance in (2.1), K^a is ensured to be positive definite. Second, for a given manifold \mathcal{M} , the geodesic distance can be specified in many ways through different Riemannian metrics on \mathcal{M} (section 3.1). According to Lemma 3.6, all these geodesic distances are equivalent to each other and the Euclidean distance on \mathbb{R}^D . Therefore, by using the Euclidean distance, we bypass the need to estimate geodesic distance, but still reflect the geometric structure of the observed predictors in terms of pairwise distances.

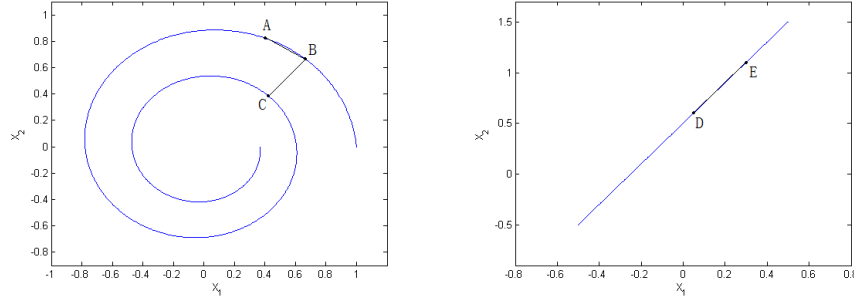
We provide heuristic explanations on why the rate can adapt to the predictor manifold through two observations. The first focuses on the possibility of obtaining an intrinsic rate for the regression problem (1.1) per se. Although the ambient space is \mathbb{R}^D , the support \mathcal{M} of the predictor X is a d dimension submanifold of \mathbb{R}^D . As a result, the GP prior specified in section 2.1 has all probability mass on the functions supported on this support, leading the posterior contraction rate to entirely depend on the evaluations of f on \mathcal{M} . More specifically, the posterior contraction rate is lower bounded by any sequence $\{\epsilon_n : n \geq 1\}$ such that

$$\Pi(d(f, f_0) > \epsilon_n | X^n) \rightarrow 0, \quad n \rightarrow \infty,$$

where $\Pi(A|X^n)$ is the posterior probability of A and $d^2(f, f_0) = (1/n) \sum_{i=1}^n (f(x_i) - f_0(x_i))^2$ under fixed design or $d^2(f, f_0) = \int_{\mathcal{M}} (f(x) - f_0(x))^2 G(dx)$ under random design, with G the marginal distribution for predictor X . Hence, $d(f, f_0)$ measures the discrepancy between f and the truth f_0 , and only depends on the evaluation of f on \mathcal{M} . Therefore, in a prediction perspective, we only need to fit and infer f on \mathcal{M} . Intuitively, we can consider a special case when the points on manifold \mathcal{M} have a global smooth representation $x = \phi(t)$, where $t \in \mathbb{R}^d$ is the global latent coordinate of x . Then the regression function

$$f(x) = f[\phi(t)] \triangleq h(t), \quad t \in \mathbb{R}^d, \quad (2.4)$$

is essentially a d -variate s -smooth function if ϕ is sufficiently smooth. Then estimation of f on \mathbb{R}^D boils down to estimation of h on \mathbb{R}^d and the intrinsic rate would be attainable. For the general case, we can consider parameterizing a

FIG 2. Examples of one dimensional submanifolds in \mathbb{R}^2 .

compact manifold \mathcal{M} by a finite number of local charts $\{(U_i, \phi_i) : i = 1, \dots, m\}$ and obtain (2.4) for x in each local neighborhood $U_i \subset \mathcal{M}$. However, since the parametrization μ in (2.4) is unknown or even does not exist, one possible goal is to develop methods that can adapt to low dimensional manifold structure.

This motivates the second observation on the possibility of obtaining the intrinsic rate via the ambient space GP prior specified in (2.2). With this prior, the dependence among $\{f(x_i)\}_{i=1}^n$ is entirely characterized by the covariance matrix $(K^A(x_i, x_j))_{n \times n}$, which depends on the pairwise Euclidean distance e among observed predictors $\{x_i\}_{i=1}^n$. Ideally, a distance $d_{\mathcal{M}}$ used in the covariance matrix should be an intrinsic distance, which measures the distance by traveling from one point to the other without leaving \mathcal{M} . More formally, an intrinsic distance is defined as the infimum of the length of all paths between two points. In the special case of (2.4), $d_{\mathcal{M}}(x, x')$ would be $e(\phi^{-1}(x), \phi^{-1}(x'))$ if ϕ is an isometric embedding from \mathbb{R}^d into \mathbb{R}^D . Fig. 2 also gives two simple examples where \mathcal{M} is a one dimensional submanifold in \mathbb{R}^2 . Although B and C are close in Euclidean distance, they are far away in terms of intrinsic distance, which is the length of the arc from B to C . Fortunately, Lemma 3.6 in the next section suggests that for compact submanifolds, this bad phenomenon only occurs for remote points — d and $d_{\mathcal{M}}$ will become comparable as two points move close. Moreover, as two points A and B become closer, using d to approximate the intrinsic distance $d_{\mathcal{M}}$ only introduces higher order error (see Proposition 3.5) proportional to the curvature of \mathcal{M} , which characterizes local distortion. In contrast, in the right plot is a straight segment in \mathbb{R}^2 . In this case Euclidean distance always matches the intrinsic distance and whether the \mathcal{M} itself is known would make no difference in predicting f since a straight segment is locally flat and has zero curvature.

A typical nonparametric approach estimates $f(x)$ by utilizing data at points near x , such as averaging over samples in a δ_n -ball around x , where the bandwidth δ_n decreases with sample size n . It is expected that as more observations

come in, properly shrinking δ_n could suppress both bias and variance, where the former is caused by local averaging and the latter is due to measurement error. This is only possible when f has certain smoothness such that large local fluctuations are not allowed. Therefore bandwidth tends to decrease at rate $n^{-1/(2s+d)}$ depending on the smoothness level s of f . Since the scaling parameter a in the covariance function K^a serves as an inverse bandwidth which would grow at rate $n^{1/(2s+d)}$, remote points tend to have exponentially decaying impact. As a result, one can imagine that accurate approximation of local intrinsic distance could provide good recovery of f as if we know the manifold and the associated intrinsic metric $d_{\mathcal{M}}$. Note that for manifold \mathcal{M} , the notion of “closeness” is characterized by the geodesic distances defined on \mathcal{M} . Often geodesic distances on \mathcal{M} are not uniquely determined (section 3.1). Fortunately, Lemma 3.6 implies that for compact submanifolds, all distance metrics induced by Riemannian metrics on \mathcal{M} are equivalent. Therefore we can choose any valid Riemannian metric as the base metric, which is the one induced by the ambient Euclidean metric in this paper. The following theorem is our main result which formalizes the above observations.

Theorem 2.1. *Assume that \mathcal{M} is a d -dimensional compact C^γ submanifold of R^D . For any $f_0 \in C^s(\mathcal{M})$ with $s \leq \min\{2, \gamma - 1\}$, if we specify the prior as (2.2), then (4.1) will be satisfied for ϵ_n a multiple of $n^{-s/(2s+d)}(\log n)^{\kappa_1}$ and $\bar{\epsilon}_n$ a multiple of $\epsilon_n(\log n)^{\kappa_2}$ with $\kappa_1 = (1 + d)/(2 + d/s)$ and $\kappa_2 = (1 + d)/2$. This implies that the posterior contraction rate will be at least a multiple of $n^{-s/(2s+d)}(\log n)^{d+1}$.*

The ambient space dimension D implicitly influences the rate via a multiplicative constant. This theorem suggests that the Bayesian model (2.3) can adapt to both the low dimensional manifold structure of X and the smoothness $s \leq 2$ of the regression function. The reason the near optimal rate can only be allowed for functions with smoothness $s \leq 2$ is the order of error in approximating the intrinsic distance $d_{\mathcal{M}}$ by the Euclidean distance d (Proposition 3.5). Even if the intrinsic dimensionality d is misspecified as d' , the following theorem still ensures the rate to be much better than $n^{-O(1/D)}$ when d' is not too small.

Theorem 2.2. *Assume the same conditions as in Theorem 2.1, but with the prior specified as (2.2) with $d' \neq d$ and $d' > d^2/(2s + d)$.*

1. *If $d' > d$, then the posterior contraction rate will be at least a multiple of $n^{-s/(2s+d')}(\log n)^\kappa$, where $\kappa = (1 + d)/(2 + d'/s)$;*
2. *If $\frac{d^2}{2s+d} < d' < d$, then the posterior contraction rate will be at least a multiple of $n^{-\frac{(2s+d)d'-d^2}{2(2s+d)d'}}(\log n)^\kappa$, where $\kappa = (d + d^2)/(2d' + dd'/s) + (1 + d)/2$.*

2.3. Dimensionality Reduction and Diffeomorphism Invariance

[27] and [24] initiated the area of manifold learning, which aims to design non-linear dimensionality reduction algorithms to map high dimensional data into

$$\begin{array}{ccc}
(\mathcal{M}, g_{\mathcal{M}}) & \xrightarrow{I_d} & (\mathcal{M}, \tilde{g}_{\mathcal{M}}) \\
\downarrow \Phi & & \downarrow \tilde{\Phi} \\
(\mathbb{R}^D, e) & \xrightarrow{\Psi} & (\mathbb{R}^{\tilde{d}}, \tilde{e})
\end{array}$$

FIG 3. (Communicative) diagrams explaining the relationship between original ambient space and feature space.

a low dimensional feature space under the assumption that data fall on an embedded non-linear manifold within the high dimensional ambient space. A combination of manifold learning and usual nonparametric regression leads to a two-stage approach, in which a dimensionality reduction map from the original ambient space \mathbb{R}^D to a feature space $\mathbb{R}^{\tilde{d}}$ is estimated in the first stage and a nonparametric regression analysis with low dimensional features as predictors is conducted in the second stage. As a byproduct of Theorem 2.1, we provide a theoretical justification for this two stage approach under some mild conditions.

Fig. 3 describes relationships used in formalizing this theory. The original predictor manifold \mathcal{M} sits in the ambient space \mathbb{R}^D . A Riemannian metric $g_{\mathcal{M}}$ on \mathcal{M} is induced by the embedding map Φ and the Euclidean metric e on \mathbb{R}^D . $\Psi : \mathbb{R}^D \rightarrow \mathbb{R}^{\tilde{d}}$ is a dimensionality reduction map such that the restriction $\Psi_{\mathcal{M}}$ of Ψ on the embedding image $\Phi(\mathcal{M}) \simeq \mathcal{M}$ is a diffeomorphism, which requires $\Psi_{\mathcal{M}}$ to be injective and both $\Psi_{\mathcal{M}}$ and its inverse to be smooth. The former requirement would imply $\tilde{d} \geq d$. Diffeomorphism is the least and only requirement such that both the intrinsic dimension d of predictor X and smoothness s of regression function f are invariant. Ψ will naturally induce an embedding

$$\tilde{\Phi} = \Psi \circ \Phi : (\mathcal{M}, \tilde{g}_{\mathcal{M}}) \rightarrow (\mathbb{R}^{\tilde{d}}, \tilde{e}), \quad (2.5)$$

where the new Riemannian metric $\tilde{g}_{\mathcal{M}}$ is induced by the Euclidean metric \tilde{e} of $\mathbb{R}^{\tilde{d}}$. Finally I_d is an identity map between the same set \mathcal{M} with different Riemannian metrics. Such a map Ψ could also be chosen so that the induced embedding $\tilde{\Phi}$ satisfies some good properties, such as the equivariant embedding in shape analysis [13]. Due to the dimensionality reduction, the regression function becomes

$$f(x) = f[\Psi_{\mathcal{M}}^{-1}(\tilde{x})] \triangleq \tilde{f}(\tilde{x}),$$

where \tilde{f} is a well defined function on the manifold \mathcal{M} represented in $\mathbb{R}^{\tilde{d}}$ and has the same smoothness as f . Therefore, by specifying a GP prior (2.2) directly on $\mathbb{R}^{\tilde{d}}$, we would be able to achieve a posterior contraction rate at least $n^{-s/(2s+d)}(\log n)^{d+1}$. The above heuristic can be formalized into the following theorem.

Theorem 2.3. Assume that \mathcal{M} is a d -dimensional compact C^γ submanifold of \mathbb{R}^D . Suppose that $\Psi : \mathbb{R}^D \rightarrow \mathbb{R}^{\tilde{d}}$ is an ambient space mapping (dimension

reduction) such that Ψ restricted on $\Phi(\mathcal{M})$ is a $C^{\gamma'}$ -diffeomorphism onto its image. Then by specifying the prior (2.2) with $\{\Psi(X_i)\}_{i=1}^n$ as observed predictors and Euclidean norm of $\mathbb{R}^{\tilde{d}}$ as $\|\cdot\|$ in (2.1), for any $f_0 \in C^s(\mathcal{M})$ with $s \leq \min\{2, \gamma - 1, \gamma' - 1\}$, (4.1) will be satisfied for $\epsilon_n = n^{-s/(2s+d)}(\log n)^{\kappa_1}$ and $\bar{\epsilon}_n = \epsilon_n(\log n)^{\kappa_2}$ with $\kappa_1 = (1+d)/(2+d/s)$ and $\kappa_2 = (1+d)/2$. This implies that the posterior contraction rate will be at least $\epsilon_n = n^{-s/(2s+d)}(\log n)^{d+1}$.

2.4. Measurement Error in the Predictors

In applications, predictor X_i may not exactly lie on the manifold \mathcal{M} . We assume that $X_i = X_{i0} + \epsilon_i$, where $X_{i0} \in \mathcal{M}$ falls on the manifold and $\epsilon_i \sim N_D(0, \sigma_X^2 I_D)$ are i.i.d measurement errors. In this case, choosing a linear projection map $\Psi^P \in \mathbb{R}^{\tilde{d} \times D}$ as the dimensionality reduction Ψ in the previous section can provide huge gain in terms of smoothing the data. As long as the elements of Ψ^P do not have large variations, the central limit theorem ensures that the noise part $\Psi^P \epsilon$ has order $O_p(D^{-1/2})$, where $\epsilon = (\epsilon_1, \dots, \epsilon_n) \in \mathbb{R}^{D \times n}$. It is not straightforward to deterministically specify a linear projection Ψ^P having good properties. Hence, we consider randomly generating Ψ^P by sampling the elements i.i.d from a common distribution. The following multiplier central limit theorem [32, Lemma 2.9.5] provides support.

Lemma 2.4. *Let Z_1, Z_2, \dots be i.i.d. Euclidean random vectors with $EZ_i = 0$ and $E\|Z_i\|^2 < \infty$ independent of the i.i.d. sequence ξ_1, ξ_2, \dots , with $E\xi_i = 0$ and $E\xi_i^2 = 1$. Then conditionally on Z_1, Z_2, \dots ,*

$$\sqrt{m} \sum_{j=1}^m \xi_j Z_j \rightarrow N(0, \text{cov}(Z_1)) \text{ in distribution,}$$

for almost every sequence Z_1, Z_2, \dots .

For a fixed row $\Psi_l^P = (\zeta_{l1}, \dots, \zeta_{lD})$, its i.i.d components ζ_{lj} can be viewed as ξ_j in the lemma. Denote the rows of the noise matrix ϵ by $\epsilon_{(1)}, \dots, \epsilon_{(D)}$. Viewing $\epsilon_{(j)}$ as the Z_j , by Lemma 2.4, we obtain that the new projected l th predictor vector $\Psi_l^P(X_1, \dots, X_n)^T \in \mathbb{R}^{\tilde{d}}$ has noise $\Psi_l^P \epsilon = \sum_{j=1}^D \Psi_{lj} \epsilon_j = O_p(D^{-1/2})$. Therefore, the noise in the original predictors is reduced by random projection. The question is then whether the projected predictors can be included in a GP regression without sacrificing asymptotic performance relative to using X_{i0} . The answer is the affirmative relying on Theorem 2.3 by the following argument.

Theorem 2.3 only requires that Ψ^P is a diffeomorphism when restricted on \mathcal{M} . Surprisingly, [2] (Theorem 3.1) proved more than this in the sense that for a compact d -dimensional Riemannian submanifold \mathcal{M} of \mathbb{R}^D and a column normalized random projection Ψ^P , if the projected dimension \tilde{d} is larger than $O(d\delta^{-2} \log(CD\delta^{-1}) \log(\rho^{-1}))$, where C is a positive constant depending on \mathcal{M} , then with probability at least $1 - \rho$, for every pair of points $x, y \in \mathcal{M}$, the

following holds

$$(1 - \delta) \sqrt{\frac{\tilde{d}}{D}} \leq \frac{\|\Psi^P x - \Psi^P y\|}{\|x - y\|} \leq (1 + \delta) \sqrt{\frac{\tilde{d}}{D}},$$

where $\|\cdot\|$ is the Euclidean norm in \mathbb{R}^D or $\mathbb{R}^{\tilde{d}}$. This theorem implies that Ψ^P preserve the ambient distances up to a scaling $\sqrt{\tilde{d}/D}$ on the manifold by choosing $\delta \ll 1$. In addition, this distance preservation property can also be extended to geodesic distances [2, Corollary 3.1]. Under the noised case, by normalizing the columns in Ψ^P , the noise $\Psi_l^P \epsilon$ has order $O_p(D^{-1})$, which is of higher order compare to the scaling $O(D^{-1/2})$ in this theorem. Therefore, even if noise exists, a combination of the distance preservation property with the fact that Ψ^P is a linear map implies that with large probability, Ψ^P would be a diffeomorphism when restricted on \mathcal{M} . Then Theorem 2.3 ensures that applying random projections in the first stage and plug in these projected predictors in a second state will not sacrifice anything asymptotically relative to using X_{i0} in the GP.

3. Geometric Properties

We introduce some concepts and results in differential and Riemannian geometry, which play an important role in the convergence rate. For detailed definitions and notations, the reader is referred to [9].

3.1. Riemannian Manifold

A manifold is a topological space that locally resembles Euclidean space. A d -dimensional topological manifold \mathcal{M} can be described using an atlas, where an atlas is defined as a collection $\{(U_s, \phi_s)\}$ such that $\mathcal{M} = \bigcup_s U_s$ and each chart $\phi_s : V \rightarrow U_s$ is a homeomorphism from an open subset V of d -dimensional Euclidean space to an open subset U_s of \mathcal{M} . By constructing an atlas whose transition functions $\{\tau_{s,\beta} = \phi_\beta^{-1} \circ \phi_s\}$ are C^γ differentiable, we can further introduce a differentiable structure on \mathcal{M} . With this differentiable structure, we are able to define differentiable functions and their smoothness level $s \leq \gamma$. Moreover, this additional structure allows us to extend Euclidean differential calculus to the manifold. To measure distances and angles on a manifold, the notion of Riemannian manifold is introduced. A Riemannian manifold (\mathcal{M}, g) is a differentiable manifold \mathcal{M} in which each tangent space $T_p \mathcal{M}$ is equipped with an inner product $\langle \cdot, \cdot \rangle_p = g_p(\cdot, \cdot)$ that varies smoothly in p . The family g_p of inner products is called a Riemannian metric and is denoted by g . With this Riemannian metric g , a distance $d_{\mathcal{M}}(p, q)$ between any two points $p, q \in \mathcal{M}$ can be defined as the length of the shortest path on \mathcal{M} connecting them. For a given manifold \mathcal{M} , such as the set $P(n)$ of all $n \times n$ positive symmetric matrices [19, 12], a Riemannian metric g is not uniquely determined and can be constructed

in various manners so that certain desirable properties, such as transformation or group action invariability, are valid. Although g is not uniquely determined, the smoothness of a given function f on \mathcal{M} only depends on \mathcal{M} 's differential structure instead of its Riemannian metric. Therefore, to study functions on the manifold \mathcal{M} , we could endow it with any valid Riemannian metric. Since a low dimensional manifold structure on the \mathbb{R}^D -valued predictor X is assumed in this paper, we will focus on the case in which \mathcal{M} is a submanifold of a Euclidean space.

Definition 3.1. \mathcal{M} is called a C^γ submanifold of \mathbb{R}^D if there exists an inclusion map $\Phi : \mathcal{M} \mapsto \mathbb{R}^D$, called embedding, such that Φ is a diffeomorphism between \mathcal{M} and $\Phi(\mathcal{M}) \subset \mathbb{R}^D$, which means:

- (1) Φ is injective and γ -differentiable;
- (2) The inverse $\Phi^{-1} : \Phi(\mathcal{M}) \rightarrow \mathcal{M}$ is also γ -differentiable.

A natural choice of the Riemannian metric g of \mathcal{M} is the one induced by the Euclidean metric e of \mathbb{R}^D through

$$g_p(u, v) = e_{\Phi(p)}(d\Phi_p(u), d\Phi_p(v)) = \langle d\Phi_p(u), d\Phi_p(v) \rangle_{\mathbb{R}^D}, \quad \forall u, v \in T_p\mathcal{M},$$

for any $p \in \mathcal{M}$. Under this Riemannian metric g , $d\Phi_p : T_p\mathcal{M} \mapsto d\Phi_p(T_p\mathcal{M}) \subset T_{\Phi(p)}\mathbb{R}^D$ is an isometric embedding. Nash Embedding Theorem [20] implies that any valid Riemannian metric on \mathcal{M} could be considered as being induced from a Euclidean metric of \mathbb{R}^m with a sufficiently large m . Therefore, we would use this naturally induced g as the Riemannian metric of predictor manifold \mathcal{M} when studying the posterior contraction rate of our proposed GP prior defined on this manifold. Under such choice of g , \mathcal{M} is isometrically embedded in the ambient space \mathbb{R}^D . In addition, in the rest of this paper, we will occasionally identify \mathcal{M} with $\Phi(\mathcal{M})$ when no confusion arises.

Tangent spaces and Riemannian metric can be represented in terms of local parameterizations. Let $\phi : U \mapsto \mathcal{M}$ be a chart that maps a neighborhood U of the origin in \mathbb{R}^d to a neighborhood $\phi(U)$ of $p \in \mathcal{M}$. In the case that \mathcal{M} is a C^γ submanifold of \mathbb{R}^D , ϕ itself is γ -differentiable as a function from $U \subset \mathbb{R}^d$ to \mathbb{R}^D . Given $i \in \{1, \dots, d\}$ and $q = \phi(u)$, where $u = (u_1, \dots, u_d) \in U$, define $\frac{\partial}{\partial u_i}(q)$ to be the linear functional on $C^\gamma(\mathcal{M})$ such that

$$\frac{\partial}{\partial u_i}(q)(f) = \left. \frac{d(f \circ \phi(u + te_i))}{dt} \right|_{t=0}, \quad \forall f \in C^\gamma(\mathcal{M}),$$

where the d -dimensional vector e_i has 1 in the i -th component and 0's in others. Then $\frac{\partial}{\partial u_i}(q)$ can be viewed as a tangent vector in the tangent space $T_q\mathcal{M}$. Moreover, $\{\frac{\partial}{\partial u_i}(q) : 1 \leq i \leq d\}$ forms a basis of $T_q\mathcal{M}$ so that each tangent vector $v \in T_q\mathcal{M}$ can be written as

$$v = \sum_{i=1}^d v_i \frac{\partial}{\partial u_i}(q).$$

Under this basis, the tangent space of \mathcal{M} can be identified as \mathbb{R}^d and the matrix representation of differential $d\Phi_q$ at q has a (j, i) th element given by

$$\left\{ d\Phi_q \left(\frac{\partial}{\partial u_i} \right) \right\}_j = \left. \frac{d(\Phi_j \circ \phi(u + te_i))}{dt} \right|_{t=0}, \quad i = 1, \dots, d, \quad j = 1, \dots, D,$$

where we use the notation F_j to denote the j th component of a vector-valued function F . In addition, under the same basis, the Riemannian metric g_q at q can be expressed as

$$g_q(v, w) = \sum_{i,j=1}^d v_i w_j g_{ij}^\phi(u_1, \dots, u_d),$$

where (v_1, \dots, v_d) and (w_1, \dots, w_d) are the local coordinates for $v, w \in T_q \mathcal{M}$. By the isometry assumption,

$$g_{ij}^\phi(u_1, \dots, u_d) = \langle d\Phi_q \left(\frac{\partial}{\partial u_i} \right), d\Phi_q \left(\frac{\partial}{\partial u_j} \right) \rangle_{R^D}.$$

Riemannian volume measure (form) of a region R contained in a coordinate neighborhood $\phi(U)$ is defined as

$$\text{Vol}(R) = \int_R dV(q) \triangleq \int_{\phi^{-1}(R)} \sqrt{\det(g_{ij}^\phi(u))} du_1 \dots du_d.$$

The volume of a general compact region R , which is not contained in a coordinate neighborhood, can be defined through partition of unity [9]. Vol generalizes the Lebesgue measure of Euclidean spaces and can be used to define the integral of a function $f \in C(\mathcal{M})$ as $\int_{\mathcal{M}} f(q) dV(q)$. In the special case that f is supported on a coordinate neighborhood $\phi(U)$,

$$\int_{\mathcal{M}} f(q) dV(q) = \int_U f(\phi(u)) \sqrt{\det(g_{ij}^\phi(u))} du_1 \dots du_d. \quad (3.1)$$

3.2. Exponential Map

Geodesic curves, generalizations of straight lines from Euclidean spaces to curved spaces, are defined as those curves whose tangent vectors remain parallel if they are transported and are locally the shortest path between points on the manifold. Formally, for $p \in \mathcal{M}$ and $v \in T_p \mathcal{M}$, the geodesic $\gamma(t, p, v), t > 0$, starting at p with velocity v , i.e. $\gamma(0, p, v) = p$ and $\gamma'(0, p, v) = v$, can be found as the unique solution of an ordinary differential equation. The exponential map $\mathcal{E}_p : T_p \mathcal{M} \mapsto \mathcal{M}$ is defined by $\mathcal{E}_p(v) = \gamma(1, p, v)$ for any $v \in T_p \mathcal{M}$ and $p \in \mathcal{M}$. Under this special local parameterization, calculations can be considerably simplified since quantities such as \mathcal{E}_p 's differential and Riemannian metric would have simple forms.

Although Hopf-Rinow theorem ensures that for compact manifolds the exponential map \mathcal{E}_p at any point p can be defined on the entire tangent space $T_p\mathcal{M}$, generally this map is no longer a global diffeomorphism. Therefore to ensure good properties of this exponential map, the notion of a normal neighborhood is introduced as follows.

Definition 3.2. A neighborhood V of $p \in \mathcal{M}$ is called normal if:

- (1) Every point $q \in V$ can be joined to p by a unique geodesic $\gamma(t, p, v)$, $0 \leq t \leq 1$, with $\gamma(0, p, v) = p$ and $\gamma(1, p, v) = q$;
- (2) \mathcal{E}_p is a diffeomorphism between V and a neighborhood of the origin in $T_p\mathcal{M}$.

Proposition 2.7 and 3.6 in [9] ensure that every point in \mathcal{M} has a normal neighborhood. However, if we want to study some properties that hold uniformly for all exponential maps \mathcal{E}_q with q in a small neighborhood of p , we need a notion stronger than normal neighborhood, whose existence has been established in Theorem 3.7 in [9].

Definition 3.3. A neighborhood W of $p \in \mathcal{M}$ is called uniformly normal if there exists some $\delta > 0$ such that:

- (1) For every $q \in W$, \mathcal{E}_p is defined on the δ -ball $B_\delta(0) \subset T_q\mathcal{M}$ around the origin of $T_q\mathcal{M}$. Moreover, $\mathcal{E}_p(B_\delta(0))$ is a normal neighborhood of q ;
- (2) $W \subset \mathcal{E}_p(B_\delta(0))$, which implies that W is a normal neighborhood of all its points.

Moreover, as pointed out by [11] and [33], by shrinking W and reducing δ at the same time, a special uniformly normal neighborhood can be chosen.

Proposition 3.4. For every $p \in \mathcal{M}$ there exists a neighborhood W such that:

- (1) W is a uniformly normal neighborhood of p with some $\delta > 0$;
- (2) The closure of W is contained in a strongly convex neighborhood U of p ;
- (3) The function $F(q, v) = (q, \mathcal{E}_q(v))$ is a diffeomorphism from $W_\delta = W \times B_\delta(0)$ onto its image in $\mathcal{M} \times \mathcal{M}$. Moreover, $|dF|$ is bounded away from zero on W_δ .

Here U is strongly convex if for every two points in U , the minimizing geodesic joining them also lies in U .

Throughout the rest of the paper, we will assume that the uniformly normal neighborhoods also possess the properties in the above proposition. Given a point $p \in \mathcal{M}$, we choose a uniformly normal neighborhood W of p . Let $\{e_1, \dots, e_d\}$ be an orthonormal basis of $T_p\mathcal{M}$. For each $q \in W$, we can define a set of tangent vectors $\{e_1^q, \dots, e_d^q\} \subset T_q\mathcal{M}$ by parallel transport [9]: $e_i \in T_p\mathcal{M} \mapsto e_i^{\gamma(t)} \in T_{\gamma(t)}\mathcal{M}$ from p to q along the unique minimizing geodesic $\gamma(t)$ ($0 \leq t \leq 1$) with $\gamma(0) = p, \gamma(1) = q$. Since parallel transport preserves the inner product in the sense that $g_{\gamma(t)}(v^{\gamma(t)}, w^{\gamma(t)}) = g_p(v, w), \forall v, w \in T_p\mathcal{M}$, $\{e_1^q, \dots, e_d^q\}$ forms an orthonormal basis of $T_q\mathcal{M}$. In addition, the orthonormal frame defined in this way is unique and depends smoothly on q . Therefore, we

obtain on W a system of normal coordinates at each $q \in W$, which parameterizes $x \in \mathcal{E}_q(B_\delta(0))$ by

$$x = \mathcal{E}_q\left(\sum_{i=1}^d u_i e_i^q\right) = \phi^q(u_1, \dots, u_d), \quad u = (u_1, \dots, u_d) \in B_\delta(0). \quad (3.2)$$

Such coordinates are called q -normal coordinates. The basis of $T_q\mathcal{M}$ associated with this coordinate chart $(B_\delta(0), \phi^q)$ is given by

$$\frac{\partial}{\partial u_i}(q)(f) = \frac{d(f \circ \mathcal{E}_q(te_i^q))}{dt} \Big|_{t=0} = \frac{d(f \circ \gamma(t, q, e_i^q))}{dt} \Big|_{t=0} = e_i^q(f), \quad i = 1, \dots, d.$$

Therefore $\{\frac{\partial}{\partial u_i}(q) = e_i^q : 1 \leq i \leq d\}$ forms an orthonormal basis on $T_q\mathcal{M}$. By Proposition 3.4, for each $x \in \mathcal{E}_q(B_\delta(0))$, there exists a minimizing geodesic $\gamma(t, q, v)$, $0 \leq t \leq 1$, such that $\gamma(0, q, v) = q$, $\gamma'(0, q, v) = v$ and $\gamma(1, q, v) = x$, where $v = \mathcal{E}_q^{-1}(x) = \sum_{i=1}^d u_i e_i^q \in T_q\mathcal{M}$. Hence $d_{\mathcal{M}}(q, x) = \int_0^1 |\gamma'(t, q, v)| dt = |v| = \|u\|$, i.e.

$$d_{\mathcal{M}}\left(q, \mathcal{E}_q\left(\sum_{i=1}^d u_i e_i^q\right)\right) = \|u\|, \quad \forall u \in B_{\delta_p}(0), \quad (3.3)$$

where $\|\cdot\|$ is the Euclidean norm on \mathbb{R}^d . The components $g_{ij}^q(u)$ of the Riemannian metric in q -normal coordinates satisfy $g_{ij}^q(0) = g_q(e_i^q, e_j^q) = \delta_{ij}$. The following results [11, Proposition 2.2] provide local expansions for the Riemannian metric $g_{ij}^q(u)$, the Jacobian $\sqrt{\det(g_{ij}^q(u))}$ and the distance $d_{\mathcal{M}}(q, \sum_{i=1}^d u_i e_i^q)$ in a neighborhood of p .

Proposition 3.5. *Let \mathcal{M} be a submanifold of \mathbb{R}^D which is isometrically embedded. Given a point $p \in \mathcal{M}$, let W and δ be as in Proposition 3.4, and consider for each $q \in W$ the q -normal coordinates defined above. Suppose that $x = \sum_{i=1}^d u_i e_i^q \in \mathcal{E}_q(B_\delta(0))$. Then:*

- (1) *The components $g_{ij}^q(u)$ of the metric tensor in q -normal coordinates admit the following expansion, uniformly in $q \in W$ and $x \in \mathcal{E}_q(B_\delta(0))$:*

$$g_{ij}^q(u_1, \dots, u_d) = \delta_{ij} - \frac{1}{3} \sum_{r,s=1}^d R_{irsj}^q(0) u_r u_s + O(d_{\mathcal{M}}^3(q, x)), \quad (3.4)$$

where $R_{irsj}^q(0)$ are the components of the curvature tensor at q in q -normal coordinates.

- (2) *The Jacobian $\sqrt{\det(g_{ij}^q)}(u)$ in q -normal coordinates has the following expansion, uniformly in $q \in W$ and $x \in \mathcal{E}_q(B_\delta(0))$:*

$$\sqrt{\det(g_{ij}^q)}(u_1, \dots, u_d) = 1 - \frac{1}{6} \sum_{r,s=1}^d Ric_{rs}^q(0) u_r u_s + O(d_{\mathcal{M}}^3(q, x)), \quad (3.5)$$

where $Ric_{rs}^q(0)$ are the components of the Ricci tensor at q in q -normal coordinates.

(3) There exists $C_p < \infty$ such that

$$0 \leq d_{\mathcal{M}}^2(q, x) - \|q - x\|^2 \leq C_p d_{\mathcal{M}}^4(q, x) \quad (3.6)$$

holds uniformly in $q \in W$ and $x \in \mathcal{E}_q(B_\delta(0))$.

Note that in Proposition 3.5, (3) only provides a comparison of geodesic distance and Euclidean distance in local neighborhoods. Under a stronger compactness assumption on \mathcal{M} , the following lemma offers a global comparison of these two distances.

Lemma 3.6. *Let \mathcal{M} be a connected compact submanifold of \mathbb{R}^D with a Riemannian metric g that is not necessarily induced from the Euclidean metric. Then there exist positive constants C_1 and C_2 dependent on g , such that*

$$C_1 \|x - y\| \leq d_{\mathcal{M}}(x, y) \leq C_2 \|x - y\|, \quad \forall x, y \in \mathcal{M}, \quad (3.7)$$

where $\|\cdot\|$ is the Euclidean distance in \mathbb{R}^D . Moreover, if \mathcal{M} is further assumed to be isometrically embedded, i.e. g is induced from the Euclidean metric of \mathbb{R}^D , then C_1 could be chosen to be one and $C_2 \geq 1$.

Proof. We only prove the first half of the inequality since the second half follows by a similar argument and is omitted here. Assume in the contrary that for any positive integer k , there exists (x_k, y_k) such that $\|x_k - y_k\| \geq k d_{\mathcal{M}}(x_k, y_k)$. Let $\Phi : \mathcal{M} \rightarrow \mathbb{R}^D$ be the embedding. Since \mathcal{M} is compact, $\{x_k\}$ and $\{y_k\}$ have convergent subsequences, whose notations are abused as $\{x_k\}$ and $\{y_k\}$ for simplicity. Denote the limits of these two sequences as x_0 and y_0 . By the compactness of \mathcal{M} and continuity of Φ , we know that $\Phi(\mathcal{M})$ is also compact and therefore $d_{\mathcal{M}}(x_k, y_k) \rightarrow 0$, as $k \rightarrow \infty$. This implies that $x_0 = y_0 = p$.

For each $j \in \{1, \dots, p\}$, the j th component $\Phi_j : \mathcal{M} \rightarrow \mathbb{R}$ of Φ is differentiable. Let δ_p and W_p be the δ and W specified in Proposition 3.4. Define $f(q, v) = \Phi(\pi_2(F(q, v))) = \Phi(\mathcal{E}_p(v))$, where π_2 is the projection of $\mathcal{M} \times \mathcal{M}$ on to its second component. By Proposition 3.4, f is differentiable on the compact set \bar{W}_{δ_p} and therefore for each $l \in \{1, \dots, d\}$, $\frac{\partial f}{\partial v_l}$ is uniformly bounded on \bar{W}_{δ_p} . This implies that for some constant $C > 0$, $\|x - y\| = \|f(y, \mathcal{E}_y^{-1}(x)) - f(y, \mathcal{E}_y^{-1}(y))\| \leq C \|\mathcal{E}_y^{-1}(x) - \mathcal{E}_y^{-1}(y)\| = C d_{\mathcal{M}}(x, y)$ for all $x, y \in W_p$ with $d_{\mathcal{M}}(x, y) \leq \delta_p$. Since $x_k \rightarrow p$ and $y_k \rightarrow p$, there exists an integer k_0 such that for all $k > k_0$, $x_k, y_k \in W_p$ and $d_{\mathcal{M}}(x_k, y_k) \leq \delta_p$. Therefore $\|x_k - y_k\| \leq C d_{\mathcal{M}}(x_k, y_k)$, which contradict our assumption that $\|x_k - y_k\| \geq k d_{\mathcal{M}}(x_k, y_k)$ for all k .

Consider the case when Φ is an isometric embedding. For any points $x, y \in \mathcal{M}$, we can cover the compact geodesic path $l_{x,y}$ from x to y by $\{W_{p_i} : i = 1, \dots, n\}$ associated with a finite number of points $\{p_1, \dots, p_n\} \subset \mathcal{M}$. Therefore we can divide $l_{x,y}$ into $\bigcup_{s=1}^n l(x_{s-1}, x_s)$ such that $x_0 = x$, $x_n = y$, and each segment $l(x_{s-1}, x_s)$ lies in one of the W_{p_i} 's. By Proposition 3.5 (3), for each $s \in \{1, \dots, n\}$, $d_{\mathcal{M}}(x_{s-1}, x_s) \geq \|x_{s-1} - x_s\|$. Therefore,

$$d_{\mathcal{M}}(x, y) = \sum_{s=1}^n d_{\mathcal{M}}(x_{s-1}, x_s) \geq \sum_{s=1}^n \|x_{s-1} - x_s\| \geq \|x - y\|,$$

where the last step follows from the triangle inequality. \square

The above lemma also implies that geodesic distances induced by different Riemannian metrics on \mathcal{M} are equivalent to each other.

Fix $p \in \mathcal{M}$ and let W and $\delta > 0$ be specified as in Proposition 3.4. Since \mathcal{M} is a submanifold of \mathbb{R}^D , for any point $q \in \mathcal{M}$, the exponential map $\mathcal{E}_q : B_\delta(0) \rightarrow \mathcal{M} \subset \mathbb{R}^D$ is a differentiable function between two subsets of Euclidean spaces. Here, we can choose any orthonormal basis of $T_q\mathcal{M}$ since the representations of \mathcal{E}_q under different orthonormal bases are the same up to $d \times d$ rotation matrices. Under the compactness assumption on \mathcal{M} , the following lemma, which will be applied in the proof of lemma 4.4, ensures the existence of a bound on the partial derivatives of \mathcal{E}_q 's components $\{\mathcal{E}_{q,i} : i = 1, \dots, D\}$ uniformly for all q in the δ neighborhood of p :

Lemma 3.7. *Let \mathcal{M} be a connected C^γ compact submanifold of \mathbb{R}^D with γ being ∞ or any integer greater than two. Let k be an integer such that $k \leq \gamma$. Then:*

1. *There exists a universal positive number δ_0 , such that for every $p \in \mathcal{M}$, proposition 3.4 is satisfied with some $\delta > \delta_0$ and W_p ;*
2. *With this δ_0 , for any $p \in \mathcal{M}$, mixed partial derivatives with order less than or equal to k of each component of \mathcal{E}_p are bounded in $B_{\delta_0}(0) \in T_p\mathcal{M}$ by a universal constant $C > 0$.*

Proof. Note that $\mathcal{M} = \bigcup_{p \in \mathcal{M}} W(p, \delta_p)$, where δ_p and $W(p, \delta_p)$ are the corresponding p dependent δ and open neighborhood W in proposition 3.4. By the compactness of \mathcal{M} , we can choose a finite covering $\{W(p_1, \delta_{p_1}), \dots, W(p_n, \delta_{p_n})\}$. Let $\delta_0 = \min\{\delta_{p_1}, \dots, \delta_{p_n}\}$. Then the first condition is satisfied with this δ_0 since for any $p \in \mathcal{M}$, W_p could be chosen as any $W(p_j, \delta_{p_j})$ that contains p .

Next we prove the second condition. For each j , we can define q -normal coordinates on $W(p_j, \delta_{p_j})$ as before such that the exponential map at each point $q \in W(p_j, \delta_{p_j})$ can be parameterized as (3.2). Define $F_j : W(p_j, \delta_{p_j}) \times B_{\delta_{p_j}}(0) \rightarrow \mathbb{R}^D$ by $F_j(q, u) = \mathcal{E}_q(\sum_{i=1}^d u_i e_i^q) = \phi^q(u)$. Then any order k mixed partial derivative $\frac{\partial^k \phi_j^q}{\partial u_{i_1} \dots \partial u_{i_k}}(u)$ of $F_j(q, u)$ with respect to u is continuous on the compact set $W(p_j, \delta_{p_j}) \times B_{\delta_{p_j}}(0)$. Therefore these partial derivatives are bounded uniformly in $q \in W(p_j, \delta_{p_j})$ and $u \in B_{\delta_{p_j}}(0)$. Since \mathcal{M} is covered by a finite number of sets $\{W(p_1, \delta_{p_1}), \dots, W(p_n, \delta_{p_n})\}$, the second conclusion is also true. \square

By lemma 3.7, when a compact submanifold \mathcal{M} has smoothness level greater than or equal to k , we can approximate the exponential map $\mathcal{E}_p : B_{\delta_0}(0) \subset \mathbb{R}^d \rightarrow \mathbb{R}^D$ at any point $p \in \mathcal{M}$ by a local Taylor polynomial of order k with error bound $C\delta_0^k$, where C is a universal constant that only depends on k and \mathcal{M} .

4. Posterior Contraction Rate of the GP on Manifold

In the GP prior (2.3), the covariance function $K^a : \mathcal{M} \times \mathcal{M} \rightarrow R$ is essentially defined on the submanifold \mathcal{M} . Therefore, (2.3) actually defines a GP

on \mathcal{M} and we can study its posterior contraction rate as a prior for functions on the manifold. In this section, we combine geometry properties and Bayesian nonparametric asymptotic theory to prove the theorems in section 2.

4.1. Reproducing Kernel Hilbert Space on the Manifold

Being viewed as a covariance function defined on $[0, 1]^D \times [0, 1]^D$, $K^a(\cdot, \cdot)$ corresponds to a reproducing kernel Hilbert space (RKHS) \mathbb{H}^a , which is defined as the completion of \mathcal{H} , the linear space of all functions on $[0, 1]^D$ with the following form

$$x \mapsto \sum_{i=1}^m a_i K^a(x_i, x), x \in [0, 1]^D,$$

indexed by $a_1, \dots, a_m \in \mathbb{R}$ and $x_1, \dots, x_m \in [0, 1]^D$, $m \in \mathbb{N}$, relative to the norm induced by the inner product defined through $\langle K^a(x, \cdot), K^a(y, \cdot) \rangle_{\mathbb{H}^a} = K^a(x, y)$. Similarly, $K^a(\cdot, \cdot)$ can also be viewed as a covariance function defined on $\mathcal{M} \times \mathcal{M}$, with the associated RKHS denoted by $\tilde{\mathbb{H}}^a$. Here $\tilde{\mathbb{H}}^a$ is the completion of $\tilde{\mathcal{H}}$, which is the linear space of all functions on \mathcal{M} with the following form

$$x \mapsto \sum_{i=1}^m a_i K^a(x_i, x), x \in \mathcal{M},$$

indexed by $a_1, \dots, a_m \in \mathbb{R}$ and $x_1, \dots, x_m \in \mathcal{M}$, $m \in \mathbb{N}$.

Many probabilistic properties of GPs are closely related to the RKHS associated with its covariance function. Readers can refer to [1] and [30] for introductions on RKHS theory for GPs on Euclidean spaces. In order to generalize RKHS properties in Euclidean spaces to submanifolds, we need a link to transfer the theory. The next lemma achieves this by characterizing the relationship between \mathbb{H}^a and $\tilde{\mathbb{H}}^a$.

Lemma 4.1. *For any $f \in \tilde{\mathbb{H}}^a$, there exists $g \in \mathbb{H}^a$ such that $g|_{\mathcal{M}} = f$ and $\|g\|_{\mathbb{H}^a} = \|f\|_{\tilde{\mathbb{H}}^a}$, where $g|_{\mathcal{M}}$ is the restriction of g on \mathcal{M} . Moreover, for any other $g' \in \mathbb{H}^a$ with $g'|_{\mathcal{M}} = f$, we have $\|g'\|_{\mathbb{H}^a} \geq \|f\|_{\tilde{\mathbb{H}}^a}$, which implies $\|f\|_{\tilde{\mathbb{H}}^a} = \inf_{g \in \mathbb{H}^a, g|_{\mathcal{M}} = f} \|g\|_{\mathbb{H}^a}$.*

Proof. Consider the map $\Phi : \tilde{\mathcal{H}} \rightarrow \mathcal{H}$ that maps the function

$$\sum_{i=1}^m a_i K^a(x_i, \cdot) \in \tilde{\mathcal{H}}, a_1, \dots, a_m \in \mathbb{R}, x_1, \dots, x_m \in \mathcal{M}, m \in \mathbb{N}$$

on \mathcal{M} to the function of the same form

$$\sum_{i=1}^m a_i K^a(x_i, \cdot) \in \mathcal{H},$$

but viewed as a function on $[0, 1]^D$. By definitions of RKHS norms, Φ is an isometry between $\tilde{\mathcal{H}}$ and a linear subspace of \mathcal{H} . As a result, Φ can be extended

to an isometry between $\tilde{\mathbb{H}}^a$ and a complete subspace of \mathbb{H}^a . To prove the first part of this lemma, it suffices to justify that for any $f \in \tilde{\mathbb{H}}^a$, $g = \Phi(f)|_{\mathcal{M}} = f$. Assume that the sequence $\{f_n\} \in \tilde{\mathcal{H}}$ satisfies

$$\|f_n - f\|_{\tilde{\mathbb{H}}^a} \rightarrow 0, \text{ as } n \rightarrow \infty,$$

then by the definition of Φ on $\tilde{\mathcal{H}}$, $\Phi(f_n)|_{\mathcal{M}} = f_n$. For any $x \in [0, 1]^D$, by the reproducing property and Cauchy-Schwarz inequality,

$$\begin{aligned} |\Phi(f_n)(x) - g(x)| &= |\langle K^a(x, \cdot), \Phi(f_n) - g \rangle_{\mathbb{H}^a}| \\ &\leq \sqrt{K^a(x, x)} \|\Phi(f_n) - \Phi(f)\|_{\mathbb{H}^a} \\ &= \|f_n - f\|_{\tilde{\mathbb{H}}^a} \rightarrow 0, \text{ as } n \rightarrow \infty, \end{aligned}$$

where the last step is by isometry. This indicates that g can be obtained as a point limit of $\Phi(f_n)$ on $[0, 1]^D$ and in the special case when $x \in \mathcal{M}$,

$$g(x) = \lim_{n \rightarrow \infty} \Phi(f_n)(x) = \lim_{n \rightarrow \infty} f_n(x) = f(x).$$

Denote the orthogonal complement of $\Phi(\tilde{\mathbb{H}}^a)$ in \mathbb{H}^a as $\Phi(\tilde{\mathbb{H}}^a)^\perp$. Since $(g' - g)|_{\mathcal{M}} = 0$, which means $\langle K^a(x, \cdot), g - g' \rangle_{\mathbb{H}^a} = 0$ for all $x \in \mathcal{M}$. Therefore by the previous construction, $g - g' \perp \Phi(\tilde{\mathbb{H}}^a)$, i.e. $g' - g \in \Phi(\tilde{\mathbb{H}}^a)^\perp$ and using Pythagorean theorem, we have

$$\|g'\|_{\mathbb{H}^a} = \|g\|_{\mathbb{H}^a} + \|g - g'\|_{\mathbb{H}^a} \geq \|g\|_{\mathbb{H}^a}.$$

□

This lemma implies that any element f in the RKHS $\tilde{\mathbb{H}}^a$ could be considered as the restriction of some element g in the RKHS \mathbb{H}^a . Particularly, there exists a unique such element g in \mathbb{H}^a such that the norm is preserved, i.e. $\|g\|_{\mathbb{H}^a} = \|f\|_{\tilde{\mathbb{H}}^a}$.

4.2. Background on Posterior Convergence Rate for GP

As shown in [10], in order to characterize the posterior contraction rate in a Bayesian nonparametric problem, such as density estimation, fixed/random design regression or classification, we need to verify some conditions on the prior measure Π . Specifically, we describe the sufficient conditions for randomly rescaled GP prior as (2.2) given in [31]. Let \mathcal{X} be the predictor space and f_0 be the true function $f_0 : \mathcal{X} \rightarrow \mathbb{R}$, which is the log density $\log p(x)$ in density estimation, regression function $E[Y|X]$ in regression and logistic transformed conditional probability $\text{logit}P(Y = 1|X)$ in classification. We will not consider density estimation since to specify the density by log density f_0 , we need to know the support \mathcal{M} so that e^{f_0} can be normalized to produce a valid density. Let ϵ_n and $\bar{\epsilon}_n$ be two sequences. If there exist Borel measurable subsets B_n of $C(\mathcal{X})$ and constant $K > 0$ such that for n sufficiently large,

$$\begin{aligned} P(\|W^A - f_0\|_\infty \leq \epsilon_n) &\geq e^{-n\epsilon_n^2}, \\ P(W^A \notin B_n) &\leq e^{-4n\epsilon_n^2}, \\ \log N(\bar{\epsilon}_n, B_n, \|\cdot\|_\infty) &\leq n\bar{\epsilon}_n^2, \end{aligned} \tag{4.1}$$

where $W^A \sim \Pi$ and $\|\cdot\|_\infty$ is the sup-norm on $C(\mathcal{X})$, then the posterior contraction rate would be at least $\epsilon_n \vee \bar{\epsilon}_n$. In our case, \mathcal{X} is the d -dimensional submanifold \mathcal{M} in the ambient space \mathbb{R}^D . To verify the first concentration condition, we need to give upper bounds to the so-called concentration function [31] $\phi_{f_0}^a(\epsilon)$ of the GP W^a around truth f_0 for given a and ϵ . $\phi_{f_0}^a(\epsilon)$ is composed of two terms. Both terms depend on the RKHS $\tilde{\mathbb{H}}^a$ associated with the covariance function of the GP W^a . The first term is the decentering function $\inf\{\|h\|_{\tilde{\mathbb{H}}^a}^2 : \|h - f_0\|_\infty < \epsilon\}$, where $\|\cdot\|_{\tilde{\mathbb{H}}^a}$ is the RKHS norm. This quantity measures how well the truth f_0 could be approximated by the elements in the RKHS. The second term is the negative log small ball probability $-\log P(\|W^a\|_\infty < \epsilon)$, which depends on the covering entropy $\log N(\epsilon_n, \tilde{\mathbb{H}}_1^a, \|\cdot\|_\infty)$ of the unit ball in the RKHS $\tilde{\mathbb{H}}^a$. As a result of this dependence, by applying Borell's inequality [30], the second and third conditions can often be proved as byproducts by using the conclusion on the small ball probability.

As pointed out by [31], the key to ensure the adaptability of the GP prior on Euclidean spaces is a sub-exponential type tail of its stationary covariance function's spectral density, which is true for squared exponential and Matérn class covariance functions. More specifically, a squared exponential covariance function $K_1(x, y) = \exp\{-\|x - y\|^2/2\}$ on \mathbb{R}^D has a spectral representation as

$$K_1(x, y) = \int_{\mathbb{R}^D} e^{-i(\lambda, x-y)} \mu(d\lambda),$$

where μ is its spectral measure with a sub-Gaussian tail, which is lighter than sub-exponential tail in the sense that: for any $\delta > 0$,

$$\int e^{\delta\|\lambda\|} \mu(d\lambda) < \infty. \quad (4.2)$$

For convenience, we will focus on squared exponential covariance function, since generalizations to other covariance functions with sub-exponential decaying spectral densities are possible with more elaboration.

4.3. Decentering Function

To estimate the decentering function, the key step is to construct a function $I_a(f)$ on the manifold \mathcal{M} to approximate a differentiable function f , so that the RKHS norm $\|I_a(f)\|_{\tilde{\mathbb{H}}^a}$ can be tightly upper bounded. Unlike in Euclidean spaces where functions in the RKHS \mathbb{H}^a can be represented via Fourier transformations [31], there is no general way to represent and calculate RKHS norms of functions in the RKHS $\tilde{\mathbb{H}}^a$ on manifold. Therefore in the next lemma, we provide a direct way to construct the approximation function $I_a(f)$ for any truth f via convolving f with K^a on manifold \mathcal{M} :

$$\begin{aligned} I_a(f)(x) &= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathcal{M}} K^a(x, y) f(y) dV(y) \\ &= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathcal{M}} \exp\left\{-\frac{a^2\|x - y\|^2}{2}\right\} f(y) dV(y), \quad x \in \mathcal{M}, \end{aligned} \quad (4.3)$$

where V is the Riemannian volume form of \mathcal{M} . Heuristically, for large a , the above integrand only has non-negligible value in a small neighborhood around x . Therefore we can conduct a change of variable in the above integral with transformation $\phi^x : B_\delta \rightarrow W$ defined by (3.2) in a small neighborhood W of x :

$$\begin{aligned} I_a(f)(x) &= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathbb{R}^d} \exp\left\{-\frac{a^2\|\phi^x(u) - \phi^x(0)\|^2}{2}\right\} f(\phi^x(u)) \sqrt{\det(g_{ij}^\phi(u))} du, \\ &\approx \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathbb{R}^d} \exp\left\{-\frac{a^2\|u\|^2}{2}\right\} f(\phi^x(u)) du, \\ &\approx f(\phi^x(0)) = f(x), \quad x \in \mathcal{M}, \end{aligned}$$

where the above approximation holds since: 1. $\phi^x(0) = x$; 2. ϕ^x preserve local distances (Proposition 3.5 (3)); 3. the Jacobian $\sqrt{\det(g_{ij}^\phi(u))}$ is close to one (Proposition 3.5 (2)). From this heuristic argument, we can see that the approximation error $\|I_a(w) - f_0\|_\infty$ is determined by two factors: the convolution error $\left| \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathbb{R}^d} \exp\left\{-\frac{a^2\|u\|^2}{2}\right\} f(\phi^x(u)) du - f(x) \right|$ and the non-flat error caused by the nonzero curvature of \mathcal{M} . Moreover, we can expand each of these errors as a polynomial of $1/a$ and call the expansion term related to $1/a^k$ as k th order error.

When \mathcal{M} is Euclidean space \mathbb{R}^d , the non-flat error is zero, and by Taylor expansion the convolution error has order s if $f_0 \in C^s(\mathbb{R}^d)$ and $s \leq 2$, where $C^s(\mathbb{R}^d)$ is the Holder class of s -smooth functions on \mathbb{R}^d . This is because the Gaussian kernel $\exp\{-\|(x-y)\|^2/2\}$ has a vanishing moment up to first order: $\int x \exp(-\|(x-y)\|^2/2) dx = 0$. Generally, the convolution error could have order up to $s+1$ if the convolution kernel K has vanishing moments up to order s , i.e. $\int x^t K(x) dx = 0, t = 1, \dots, s$. However, for general manifold \mathcal{M} with non-vanishing curvature tensor, the non-flat error always has order two (see the proof of Lemma 4.2). This implies that even though carefully chosen kernels for the covariance function can improve the convolution error to have order higher than two, the overall approximation still tends to have second order error due to the deterioration caused by the nonzero curvature of the manifold. The following lemma formalizes the above heuristic argument on the order of the approximation error by (4.3) and further provides an upper bound on the decentering function.

Lemma 4.2. *Assume that \mathcal{M} is a d -dimensional compact C^γ submanifold of \mathbb{R}^D . Let $C^s(\mathcal{M})$ be the set of all functions on \mathcal{M} with holder smoothness s . Then for any $f \in C^s(\mathcal{M})$ with $s \leq \min\{2, \gamma\}$, there exist constants $a_0 \geq 1$, $C > 0$ and $B > 0$ depending only on μ , \mathcal{M} and f such that for all $a \geq a_0$,*

$$\inf\{\|h\|_{\mathbb{H}^a}^2 : \sup_{x \in \mathcal{M}} |h(x) - f(x)| \leq Ca^{-s}\} \leq Ba^d.$$

Proof. The proof consists of two parts. In the first part, we prove that the approximation error of $I_a(f)$ can be decomposed into four terms. The first term T_1 is the convolution error defined in our previous heuristic argument. The second

term T_2 is caused by localization of the integration, which is negligible due to the exponential decaying of the squared exponential covariance function. The third and fourth terms T_3, T_4 correspond to the non-flat error, with T_3 caused by approximating the geodesic distance with Euclidean distance $||\phi^q(u) - q||^2 - ||u||^2$, and T_4 by approximating the Jacobian $|\sqrt{\det(g_{ij}^\phi(u))} - 1|$. Therefore the overall approximation error $|I_a(f)(x) - f(x)|$ has order s in the sense that for some constant $C > 0$ dependent on \mathcal{M} and f :

$$\sup_{x \in \mathcal{M}} |I_a(f)(x) - f(x)| \leq Ca^{-s}, \quad s \leq \min\{2, \gamma\}. \quad (4.4)$$

In the second part, we prove that $I_a(f)$ belongs to $\tilde{\mathbb{H}}^a$ and has a squared RKHS norm:

$$||I_a(f)||_{\tilde{\mathbb{H}}^a}^2 \leq Ba^d,$$

where B is a positive constant not dependent on a .

Step 1 (Estimation of the approximation error): This part follows similar ideas as in the proof of Theorem 1 in [33], where they have shown that (4.4) holds for $s \leq 1$. Our proof generalizes their results to $s \leq 2$ and therefore needs more careful estimations.

By Proposition 3.5, for each $p \in \mathcal{M}$, there exists a neighborhood W_p and an associated δ_p satisfying the two conditions in Proposition 3.4 and equations (3.4)-(3.6). By compactness, \mathcal{M} can be covered by $\cup_{p \in \mathcal{P}} W_p$ for a finite subset \mathcal{P} of \mathcal{M} . Then $\sup_{x \in \mathcal{M}} |I_a(f)(x) - f(x)| = \sup_{p \in \mathcal{P}} \{\sup_{x \in W_p} |I_a(f)(x) - f(x)|\}$. Let $\delta^* = \min_{p \in \mathcal{P}} \{\min\{\delta_p, 1/\sqrt{2C_p}\}\} > 0$, where C_p is defined as in equation (3.6). Choose $a_0 \geq 1$ sufficiently large such that $C_0\sqrt{(2d+8)\log a_0/a_0} < \delta^*$, where C_0 is the C_2 in Lemma 3.6.

Let $q \in W_p$ and $a > a_0$. Define $B_a^q = \{x \in \mathcal{M} : d_{\mathcal{M}}(q, x) < C_0\sqrt{(2d+8)\log a/a}\}$. Combining equation (3.3) and the fact that \mathcal{E}_q is a diffeomorphism on $B_{\delta^*}(0)$,

$$B_a^q = \left\{ \mathcal{E}_q \left(\sum_{i=1}^d u_i e_i^q \right) : u \in \tilde{B}_a \right\} \subset \mathcal{E}_q(B_{\delta^*}(0)),$$

where $\tilde{B}_a = \{u : ||u|| < C_0\sqrt{(2d+8)\log a/a}\} \subset B_{\delta^*}(0)$.

Denote $\phi^q(u) = \mathcal{E}_q(\sum_{i=1}^d u_i e_i^q)$. Then $B_a^q = \phi^q(\tilde{B}_a)$. By definition (3.1),

$$\begin{aligned} & \left(\frac{a}{\sqrt{2\pi}} \right)^d \int_{B_a^q} K^a(x, y) f(y) dV(y) \\ &= \left(\frac{a}{\sqrt{2\pi}} \right)^d \int_{\tilde{B}_a} \exp \left\{ - \frac{a^2 ||q - \phi^q(u)||^2}{2} \right\} f(\phi^q(u)) \sqrt{\det(g_{ij}^q)(u)} du. \end{aligned}$$

Therefore, by (4.3) we have the following decomposition:

$$I_a(f)(q) - f(q) = T_1 + T_2 + T_3 + T_4,$$

where

$$\begin{aligned}
T_1 &= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\tilde{B}_a} \exp\left\{-\frac{a^2\|u\|^2}{2}\right\} [f(\phi^q(u)) - f(\phi^q(0))] du \\
T_2 &= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathcal{M} \setminus B_a^q} K^a(q, y) f(y) dV(y) - \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathbb{R}^d \setminus \tilde{B}_a} \exp\left\{-\frac{a^2\|u\|^2}{2}\right\} f(q) du, \\
T_3 &= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\tilde{B}_a} \left\{ \exp\left\{-\frac{a^2\|q - \phi^q(u)\|^2}{2}\right\} - \exp\left\{-\frac{a^2\|u\|^2}{2}\right\} \right\} f(\phi^q(u)) du, \\
T_4 &= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\tilde{B}_a} \exp\left\{-\frac{a^2\|q - \phi^q(u)\|^2}{2}\right\} f(\phi^q(u)) (\sqrt{\det(g_{ij}^q)}(u) - 1) du.
\end{aligned}$$

Step 1.1 (Estimation of T_1): Let $g = f \circ \phi^q$. Since $f \in C^s(\mathcal{M})$ and $(\phi^q, B_{\delta^*}(0))$ is a C^γ coordinate chart, we have $g \in C^s(\mathbb{R}^d)$ and therefore

$$g(u) - g(0) = \begin{cases} R(u, s), & \text{if } 0 < s \leq \min\{1, \gamma\}, \\ \sum_{i=1}^d \frac{\partial g}{\partial u_i}(0) u_i + R(u, s), & \text{if } 1 < s \leq \min\{2, \gamma\}, \end{cases}$$

where the remainder term $|R(u, s)| \leq C_1 \|u\|^s$ for all $0 < s \leq \min\{2, \gamma\}$. Since \tilde{B}_a is symmetric,

$$\int_{\tilde{B}_a} \exp\left\{-\frac{a^2\|u\|^2}{2}\right\} u_i du = 0, \quad i = 1, \dots, d,$$

and therefore

$$|T_1| \leq C_1 \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\tilde{B}_a} \exp\left\{-\frac{a^2\|u\|^2}{2}\right\} \|u\|^s du = C_2 a^{-s}.$$

Step 1.2 (Estimation of T_2): Denote $T_2 = S_1 + S_2$ where S_1 and S_2 are the first term and second term of T_2 , respectively. By Lemma 3.6, for $y \in \mathcal{M} \setminus B_a^q$, $\|q - y\| \geq d_{\mathcal{M}}(q, y)/C_0 \geq \sqrt{(2d+8) \log a/a}$. Therefore,

$$\begin{aligned}
|S_1| &= \left| \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathcal{M} \setminus B_a^q} \exp\left\{-\frac{a^2\|q - y\|^2}{2}\right\} f(y) dV(y) \right| \\
&\leq \|f\|_\infty \text{Vol}(\mathcal{M}) \left(\frac{a}{\sqrt{2\pi}}\right)^d \exp\left\{-\frac{(2d+8) \log a}{2}\right\} \\
&= C_3 a^{-4} \leq C_3 a^{-s}.
\end{aligned}$$

As for S_2 , we have

$$\begin{aligned}
|S_2| &\leq \|f\|_\infty \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\|u\| \geq C_0 \sqrt{(2d+8) \log a/a}} \exp\left\{-\frac{a^2\|u\|^2}{2}\right\} du \\
&\leq \|f\|_\infty \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathbb{R}^d} \exp\left\{-\frac{C_0^2(2d+8) \log a}{4}\right\} \exp\left\{-\frac{a^2\|u\|^2}{4}\right\} du \\
&= C_4 a^{-C_0^2(d/2+2)} \leq C_4 a^{-s},
\end{aligned}$$

since $d \geq 1$, $C_0 \geq 1$ and $a \geq a_0 \geq 1$.

Combining the above inequalities for S_1 and S_2 , we obtain

$$|T_2| \leq (C_3 + C_4)a^{-s} = C_5a^{-s}.$$

Step 1.3 (Estimation of T_3): By equation (3.6) in Proposition 3.5 and equation (3.3), we have

$$||u|^2 - |q - \phi^q(u)|^2| = |d_{\mathcal{M}}^2(q, \phi^q(u)) - |q - \phi^q(u)|^2| \leq C_p d_{\mathcal{M}}^4(q, \phi^q(u)) = C_p ||u||^4. \quad (4.5)$$

Therefore by using the inequality $|e^{-a} - e^{-b}| \leq |a - b| \max\{e^{-a}, e^{-b}\}$ for $a, b > 0$, we have

$$|T_3| \leq \|f\|_{\infty} \left(\frac{a}{\sqrt{2\pi}} \right)^d \int_{\tilde{B}_a} \max \left\{ \exp \left\{ -\frac{a^2 ||q - \phi^q(u)||^2}{2} \right\}, \exp \left\{ -\frac{a^2 ||u||^2}{2} \right\} \right\} \frac{a^2 ||u||^4}{2} du.$$

By equation (4.5) and the fact that $u \in \tilde{B}_a$, $||u||^2 \leq (\delta^*)^2 \leq 1/(2C_p)$ and hence

$$||u||^2 - |q - \phi^q(u)|^2 \leq \frac{1}{2} ||u||^2, \quad |q - \phi^q(u)|^2 \geq \frac{1}{2} ||u||^2. \quad (4.6)$$

Therefore

$$|T_3| \leq \|f\|_{\infty} \left(\frac{a}{\sqrt{2\pi}} \right)^d \int_{\tilde{B}_a} \exp \left\{ -\frac{a^2 ||u||^2}{4} \right\} \frac{a^2 ||u||^4}{2} du = C_6 a^{-2} \leq C_6 a^{-s},$$

since $a \geq a_0 \geq 1$.

Step 1.4 (Estimation of T_4): By equation (3.5) in Proposition 3.5, there exists a constant C_7 depending on the Ricci tensor of the manifold \mathcal{M} , such that

$$|\sqrt{\det(g_{ij}^q)}(u) - 1| \leq C_7 ||u||^2.$$

Therefore, by applying equation (4.6) again, we obtain

$$|T_4| \leq C_4 \|f\|_{\infty} \left(\frac{a}{\sqrt{2\pi}} \right)^d \int_{\tilde{B}_a} \exp \left\{ -\frac{a^2 ||u||^2}{4} \right\} ||u||^2 du = C_8 a^{-2} \leq C_8 a^{-s}.$$

Combining the above estimates for T_1 , T_2 , T_3 and T_4 , we have

$$\sup_{x \in \mathcal{M}} |I_a(f)(q)(x) - f(q)(x)| \leq (C_2 + C_3 + C_6 + C_8)a^{-s} = Ca^{-s}.$$

Step 2 (Estimation of the RKHS norm): Since $\langle K^a(x, \cdot), K^a(y, \cdot) \rangle_{\tilde{\mathbb{H}}^a} = K^a(x, y)$, we have

$$\begin{aligned} ||I_a(f)||_{\tilde{\mathbb{H}}^a} &= \left(\frac{a}{\sqrt{2\pi}} \right)^{2d} \int_{\mathcal{M}} \int_{\mathcal{M}} K^a(x, y) f(x) f(y) dV(x) dV(y) \\ &\leq \|f\|_{\infty}^2 \left(\frac{a}{\sqrt{2\pi}} \right)^d \int_{\mathcal{M}} dV(x) \left(\frac{a}{\sqrt{2\pi}} \right)^d \int_{\mathcal{M}} K^a(x, y) dV(y). \end{aligned}$$

Applying the results of the first part to function $f \equiv 1$, we have

$$\left| \left(\frac{a}{\sqrt{2\pi}} \right)^d \int_{\mathcal{M}} K^a(x, y) dV(y) - 1 \right| \leq C a^{-2} \leq C,$$

since $a \geq a_0 \geq 1$. Therefore,

$$\|I_a(f)\|_{\tilde{\mathbb{H}}^a} \leq (1+C)\|f\|_{\infty}^2 \left(\frac{a}{\sqrt{2\pi}} \right)^d \text{Vol}(\mathcal{M}) = B a^d.$$

□

4.4. Centered Small Ball Probability

As indicated by the proof of Lemma 4.6 in [31], to obtain an upper bound on $-\log P(\|W^a\|_{\infty} < \epsilon)$, we need to provide an upper bound for the covering entropy $\log N(\epsilon, \tilde{\mathbb{H}}_1^a, \|\cdot\|_{\infty})$ of the unit ball in the RKHS $\tilde{\mathbb{H}}^a$ on the submanifold \mathcal{M} . Following the discussion in section 4.1, we want to link $\tilde{\mathbb{H}}^a$ to \mathbb{H}^a , the associated RKHS defined on the ambient space \mathbb{R}^D . Therefore, we need a lemma to characterize the space \mathbb{H}^a [31, Lemma 4.1].

Lemma 4.3. \mathbb{H}^a is the set of real parts of the functions

$$x \mapsto \int e^{i(\lambda, x)} \psi(\lambda) \mu_a(d\lambda),$$

when ψ runs through the complex Hilbert space $L_2(\mu_a)$. Moreover, the RKHS norm of the above function is $\|\psi\|_{L_2(\mu_a)}$, where μ_a is the spectral measure of the covariance function K^a .

Based on this representation of \mathbb{H}^a on \mathbb{R}^D , [31] proved an upper bound $K a^D (\log \frac{1}{\epsilon})^{D+1}$ for $\log N(\epsilon, \tilde{\mathbb{H}}_1^a, \|\cdot\|_{\infty})$ through constructing an ϵ -covering set composed of piecewise polynomials. However, there is no straightforward generalization of their scheme from Euclidean spaces to manifolds. The following lemma provides an upper bound for the covering entropy of $\tilde{\mathbb{H}}_1^a$, where the D in the upper bounds for \mathbb{H}_1^a is reduced to d . The main novelty in our proof is the construction of an ϵ -covering set composed of piecewise transformed polynomials (4.12) via analytically extending the truncated Taylor polynomial approximations (4.9) of the elements in $\tilde{\mathbb{H}}_1^a$. As the proof indicates, the d in a^d relates to the covering dimension d of \mathcal{M} , i.e. the ϵ -covering number $N(\epsilon, \mathcal{M}, \epsilon)$ of \mathcal{M} is proportional to $1/\epsilon^d$. The d in $(\log \frac{1}{\epsilon})^{d+1}$ relates to the order of the number k^d of coefficients in piecewise transformed polynomials of degree k in d variables.

Lemma 4.4. Assume that \mathcal{M} is a d -dimensional C^γ compact submanifold of \mathbb{R}^D with $\gamma \geq 2$. Then for squared exponential covariance function K^a , there exists a constant K depending only on d , D and \mathcal{M} , such that for $\epsilon < 1/2$ and $a > \max\{a_0, \epsilon^{-1/(\gamma-1)}\}$, where δ_0 is defined in Lemma 3.7 and a_0 is a universal constant,

$$\log N(\epsilon, \tilde{\mathbb{H}}_1^a, \|\cdot\|_{\infty}) \leq K a^d \left(\log \frac{1}{\epsilon} \right)^{d+1}.$$

Proof. By Lemma 4.1 and Lemma 4.3, a typical element of $\tilde{\mathbb{H}}^a$ can be written as the real part of the function

$$h_\psi(x) = \int e^{i(\lambda, x)} \psi(\lambda) \mu_a(d\lambda), \text{ for } x \in \mathcal{M}$$

for $\psi : \mathbb{R}^D \rightarrow \mathbb{C}$ a function with $\int |\psi|^2 \mu_a(d\lambda) \leq 1$. This function can be extended to \mathbb{R}^D by allowing $x \in \mathbb{R}^D$. For any given point $p \in \mathcal{M}$, by (4.3), we have a local coordinate $\phi^p : B_{\delta_0}(0) \subset \mathbb{R}^d \rightarrow \mathbb{R}^D$ induced by the exponential map \mathcal{E}_p . Therefore, for $x \in \phi_p(B_{\delta_0}(0))$, $h_\psi(x)$ can be written in local q -normal coordinates as

$$h_{\psi,p}(u) = h_\psi(\phi^p(u)) = \int e^{i(\lambda, \phi^p(u))} \psi(\lambda) \mu_a(d\lambda), \quad u \in B_{\delta_0}(0). \quad (4.7)$$

Similar to the idea in the proof of Lemma 4.5 in [31], we want to extend the function $h_{\psi,p}$ to an analytical function $z \mapsto \int e^{i(\lambda, \phi^p(z))} \psi(\lambda) \mu_a(d\lambda)$ on the set $\Omega = \{z \in \mathbb{C}^d : \|\operatorname{Re} z\| < \delta_0, \|\operatorname{Im} z\| < \rho/a\}$ for some $\rho > 0$. Then we can obtain upper bounds on the mixed partial derivatives of the analytic extension $h_{\psi,p}$ via Cauchy formula, and finally construct an ϵ -covering set of $\tilde{\mathbb{H}}_1^a$ by piecewise polynomials defined on \mathcal{M} . Unfortunately, this analytical extension is impossible unless $\phi^p(u)$ is a polynomial. This motivates us to approximate $\phi^p(u)$ by its γ th order Taylor polynomial $P_{p,\gamma}(u)$. More specifically, by Lemma 4.7 and the discussion after Lemma 3.7, the error caused by approximating $\phi^p(u)$ by $P_{p,\gamma}(u)$ is

$$|h_\psi(\phi^p(u)) - h_\psi(P_{p,\gamma}(u))| \leq a \|\phi^p(u) - P_{p,\gamma}(u)\| \leq Ca \|u\|^\gamma. \quad (4.8)$$

For notation simplicity, fix p as a center and denote the function $h_\psi(P_{p,\gamma}(u))$ by $r(u)$ for $u \in B_{\delta_0}$. Since $P_{p,\gamma}(u)$ is a polynomial of degree γ , view the function r as a function of argument u ranging over the product of the imaginary axes in \mathbb{C}^d , we can extend

$$r(u) = \int e^{i(\lambda, P_{p,\gamma}(u))} \psi(\lambda) \mu_a(d\lambda), \quad u \in B_{\delta_0}(0) \quad (4.9)$$

to an analytical function $z \mapsto \int e^{i(\lambda, P_{p,\gamma}(z))} \psi(\lambda) \mu_a(d\lambda)$ on the set $\Omega = \{z \in \mathbb{C}^d : \|\operatorname{Re} z\| < \delta_0, \|\operatorname{Im} z\| < \rho/a\}$ for some $\rho > 0$ sufficiently small determined by the $\delta < 1/2$ in (4.2). Moreover, by Cauchy-Schwarz inequality, $|r(z)| \leq C$ for $z \in \Omega$ and $C^2 = \int e^{\delta \|\lambda\|} \mu(d\lambda)$. Therefore, by Cauchy formula, with D^n denoting the partial derivative of orders $n = (n_1, \dots, n_d)$ and $n! = n_1! \cdots n_d!$, we have the following bound for partial derivatives of r at any $u \in B_{\delta_0}(0)$,

$$\left| \frac{D^n r(u)}{n!} \right| \leq \frac{C}{R^n}, \quad (4.10)$$

where $R = \rho/(a\sqrt{d})$. Based on the inequalities (4.8) and (4.10), we can construct an ϵ -covering set of $\tilde{\mathbb{H}}_1^a$ as follows.

Set $a_0 = \rho/(2\delta_0\sqrt{d})$, then $R < 2\delta_0$. Since $\mathcal{M} \subset [0, 1]^D$, with C_2 defined in Lemma 3.6, let $\{p_1, \dots, p_m\}$ be an $R/(2C_2)$ -net in \mathcal{M} for the Euclidean distance, and let $\mathcal{M} = \bigcup_i B_i$ be a partition of \mathcal{M} in sets B_1, \dots, B_m obtaining by assigning every $x \in \mathcal{M}$ to the closest $p_i \in \{p_1, \dots, p_m\}$. By (3.3) and Lemma 3.6

$$|(\phi^{p_i})^{-1}(x)| < C_2 \frac{R}{2C_2} = \frac{R}{2} < \delta_0, \quad (4.11)$$

where ϕ_{p_i} is the local normal coordinate chart at p_i . Therefore, we can consider the piecewise transformed polynomials $P = \sum_{i=1}^m P_{i,a_i} 1_{B_i}$, with

$$P_{i,a_i}(x) = \sum_{n, \leq k} a_{i,n} [(\phi^{p_i})^{-1}(x)]^n, \quad x \in \phi^{p_i}(B_{\delta_0}(0)). \quad (4.12)$$

Here the sum ranges over all multi-index vectors $n = (n_1, \dots, n_d) \in (\mathbb{N} \cup \{0\})^d$ with $n_{\cdot} = n_1 + \dots + n_d \leq k$. Moreover, for $y = (y_1, \dots, y_d) \in \mathbb{R}^d$, the notation y^n used above is short for $y_1^{n_1} y_2^{n_2} \dots y_d^{n_d}$. We obtain a finite set of functions by discretizing the coefficients $a_{i,n}$ for each i and n over a grid of meshwidth ϵ/R^n -net in the interval $[-C/R^n, C/R^n]$ (by (4.10)). The log cardinality of this set is bounded by

$$\log \left(\prod_i \prod_{n: n_{\cdot} \leq k} \#a_{i,n} \right) \leq m \log \left(\prod_{n: n_{\cdot} \leq k} \frac{2C/R^n}{\epsilon/R^n} \right) \leq mk^d \log \left(\frac{2C}{\epsilon} \right).$$

Since $R = \rho/(a\sqrt{d})$, we can choose $m = N(\mathcal{M}, \|\cdot\|, \rho/(2C_0 a d^{1/2})) \simeq a^d$. To complete the proof, it suffices to show that for k of order $\log(1/\epsilon)$, the resulting set of functions is a $K\epsilon$ -net for constant K depending only on μ .

For any function $f \in \tilde{\mathbb{H}}_1^a$, by Lemma 4.1, we can find a $g \in \tilde{\mathbb{H}}_1^a$ such that $g|_{\mathcal{M}} = f$. Assume that r_g (the subscript g indicates the dependence on g) is the local polynomial approximation for g defined as (4.9). Then we have a partial derivative bound on r_g as:

$$\left| \frac{D^n r_g(p_i)}{n!} \right| \leq \frac{C}{R^n}.$$

Therefore there exists a universal constant K and appropriately chosen a_i in (4.12), such that for any $z \in B_i \subset \mathcal{M}$,

$$\left| \sum_{n, > k} \frac{D^n r_g(p_i)}{n!} (z - p_i)^n \right| \leq \sum_{n, > k} \frac{C}{R^n} (R/2)^n \leq C \sum_{l=k+1}^{\infty} \frac{l^{d-1}}{2^l} \leq KC \left(\frac{2}{3} \right)^k,$$

$$\left| \sum_{n, \leq k} \frac{D^n r_g(p_i)}{n!} (z - p_i)^n - P_{i,a_i}(z) \right| \leq \sum_{n, \leq k} \frac{\epsilon}{R^n} (R/2)^n \leq \sum_{l=1}^k \frac{l^{d-1}}{2^l} \epsilon \leq K\epsilon.$$

Moreover, by (4.8) and (4.11),

$$|g(z) - r_g(z)| \leq Ca \|(\phi^{p_i})^{-1}(z)\|^\gamma \leq aR^\gamma \leq Ka^{-(\gamma-1)} < K\epsilon,$$

where the last step follows by the condition on a .

Consequently, we obtain

$$|f(z) - P_{i,n_i}(z)| = |g(z) - P_{i,n_i}(z)| \leq |g(z) - r_g(z)| + |r_g(z) - P_{i,n_i}(z)| \leq KC \left(\frac{2}{3}\right)^k + 2K\epsilon.$$

This suggests that the piecewise polynomials form a $3K\epsilon$ -net for k sufficiently large so that $(2/3)^k$ is smaller than $K\epsilon$. \square

Similar to Lemma 4.6 in [31], Lemma 4.4 implies an upper bound on $-\log P(\|W^a\|_\infty < \epsilon)$.

Lemma 4.5. *Assume that \mathcal{M} is a d -dimensional compact C^γ submanifold of \mathbb{R}^D with $\gamma \geq 2$. If K^a is the squared exponential covariance function with inverse bandwidth a , then for some $a_0 > 0$, there exist constants C and ϵ_0 that only depend on a_0 , μ , d , D and \mathcal{M} , such that, for $a \geq \max\{a_0, \epsilon^{-1/(\gamma-1)}\}$ and $\epsilon < \epsilon_0$,*

$$-\log P\left(\sup_{x \in \mathcal{M}} |W_x^a| \leq \epsilon\right) \leq Ca^d \left(\log \frac{a}{\epsilon}\right)^{d+1}.$$

Before proving Theorem 2.1, we need another two technical lemmas for preparations, which are the analogues of Lemma 4.7 and 4.8 in [31] for RKHS on Euclidean spaces.

Lemma 4.6. *For squared exponential covariance function, if $a \leq b$, then $\sqrt{a}\tilde{\mathbb{H}}_1^a \subset \sqrt{b}\mathbb{H}_1^b$.*

Proof. For any $f \in \sqrt{a}\tilde{\mathbb{H}}_1^a$, by Lemma 4.1, there exists $g \in \sqrt{a}\mathbb{H}_1^a$ such that $g|_{\mathcal{M}} = f$. By Lemma 4.7 in [31], $\sqrt{a}\mathbb{H}_1^a \subset \sqrt{b}\mathbb{H}_1^b$, so $g \in \sqrt{b}\mathbb{H}_1^b$. Again by Lemma 4.1, since $g|_{\mathcal{M}} = f$, $\|f\|_{\tilde{\mathbb{H}}^b} \leq \|g\|_{\mathbb{H}^b} \leq \sqrt{b}$, implying that $f \in \sqrt{b}\tilde{\mathbb{H}}_1^b$. \square

Lemma 4.7. *Any $h \in \tilde{\mathbb{H}}_1^a$ satisfies $|h(x)| \leq 1$ and $|h(x) - h(x')| \leq a\|x - x'\|\tau$ for any $x, x' \in \mathcal{M}$, where $\tau^2 = \int \|\lambda\|^2 d\mu(\lambda)$.*

Proof. By the reproducing property and Cauchy-Schwarz inequality

$$\begin{aligned} |h(x)| &= |\langle h, K^a(x, \cdot) \rangle_{\tilde{\mathbb{H}}^a}| \leq \|K^a(x, \cdot)\|_{\tilde{\mathbb{H}}^a} = 1 \\ |h(x) - h(x')| &= |\langle h, K^a(x, \cdot) - K^a(x', \cdot) \rangle_{\tilde{\mathbb{H}}^a}| \\ &\leq \|K^a(x, \cdot) - K^a(x', \cdot)\|_{\tilde{\mathbb{H}}^a} \\ &= \sqrt{2(1 - K^a(x, x'))}. \end{aligned}$$

By the spectral representation $K(x, x') = \int e^{i(\lambda, t)} \mu_a(d\lambda)$ and the fact that μ_a is symmetric,

$$\begin{aligned} 2(1 - K^a(x, x')) &= 2 \int (1 + i(\lambda, x - x') - e^{i(\lambda, x - x')}) \mu_a(d\lambda) \\ &\leq \|x - x'\|^2 \int \|\lambda\|^2 \mu_a(d\lambda) \\ &= a^2 \|x - x'\|^2 \int \|\lambda\|^2 \mu(d\lambda). \end{aligned}$$

\square

4.5. Posterior Contraction Rate of GP on Manifold

We provide proofs for Theorem 2.1 and Theorem 2.2.

Proof of Theorem 2.1. Define centered and decentered concentration functions of the process $W^a = (W_{ax} : x \in \mathcal{M})$ by

$$\begin{aligned}\phi_0^a(\epsilon) &= -\log P(|W^a|_\infty \leq \epsilon), \\ \phi_{f_0}^a(\epsilon) &= \inf_{h \in \tilde{\mathbb{H}}^a : \|h - f_0\|_\infty \leq \epsilon} \|h\|_{\tilde{\mathbb{H}}^a}^2 - \log P(|W^a|_\infty \leq \epsilon),\end{aligned}$$

where $|h|_\infty = \sup_{x \in \mathcal{M}} |f(x)|$ is the sup norm on the manifold \mathcal{M} . Then $P(|W^a|_\infty \leq \epsilon) = \exp(-\phi_0^a(\epsilon))$ by definition. Moreover, by the results in [14],

$$P(\|W^a - f_0\|_\infty \leq 2\epsilon) \geq e^{-\phi_{f_0}^a(\epsilon)}. \quad (4.13)$$

Suppose that $f_0 \in C^s(\mathcal{M})$ for some $s \leq \min\{2, \gamma - 1\}$. By Lemma 4.5 and Lemma 4.2, for $a > a_0$ and $\epsilon > C \max\{a^{-(\gamma-1)}, a^{-s}\} = Ca^{-s}$,

$$\phi_{f_0}^s(\epsilon) \leq Da^d + C_4 a^d \left(\log \frac{a}{\epsilon} \right)^{1+d} \leq K_1 a^d \left(\log \frac{a}{\epsilon} \right)^{1+d}.$$

Since A^d has a Gamma prior, there exists $p, C_1, C_2 > 0$, such that $C_1 a^p \exp(-D_2 a^d) \leq g(a) \leq C_2 a^p \exp(-D_2 a^d)$. Therefore by equation (4.13),

$$\begin{aligned}P(\|W^A - f_0\|_\infty \leq 2\epsilon) &\geq P(\|W^A - f_0\|_\infty \leq 2\epsilon, A \in [(C/\epsilon)^{1/s}, 2(C/\epsilon)^{1/s}]) \\ &\geq \int_{(C/\epsilon)^{1/s}}^{2(C/\epsilon)^{1/s}} e^{-\phi_{f_0}^s(\epsilon)} g(a) da \\ &\geq C_1 e^{-K_2(1/\epsilon)^{d/s}(\log(1/\epsilon))^{1+d}} \left(\frac{C}{\epsilon} \right)^{p/s} \left(\frac{C}{\epsilon} \right)^{1/s}.\end{aligned}$$

Therefore,

$$P(\|W^A - f_0\|_\infty \leq \epsilon_n) \geq \exp(-n\epsilon_n^2),$$

for ϵ_n a large multiple of $n^{-s/(2s+d)}(\log n)^{\kappa_1}$ with $\kappa_1 = (1+d)/(2+d/s)$ and sufficiently large n .

Similar to the proof of Theorem 3.1 of [31], by Lemma 4.6,

$$B_{M,r,\delta,\epsilon} = \left(M \sqrt{\frac{r}{\delta}} \tilde{\mathbb{H}}_1^r + \epsilon \mathbb{B}_1 \right) \cup \left(\bigcup_{a < \delta} (M \tilde{\mathbb{H}}_1^a) + \epsilon \mathbb{B}_1 \right),$$

with \mathbb{B}_1 the unit ball of $C(\mathcal{M})$, contains the set $M \tilde{\mathbb{H}}_1^a + \epsilon \mathbb{B}_1$ for any $a \in [\delta, r]$. Furthermore, if

$$M \geq 4\sqrt{\phi_0^r(\epsilon)} \quad \text{and} \quad e^{-\phi_0^r(\epsilon)} < 1/4, \quad (4.14)$$

then

$$P(W^A \notin B) \leq \frac{2C_2 r^{p-d+1} e^{-D_2 r^d}}{D_2 d} + e^{-M^2/8}. \quad (4.15)$$

By Lemma 4.5, equation (4.14) is satisfied if

$$M^2 \geq 16C_4r^d(\log(r/\epsilon))^{1+d}, \quad r > 1, \quad \epsilon < \epsilon_1,$$

for some fixed $\epsilon_1 > 0$. Therefore

$$P(W^A \notin B) \leq \exp(-C_0n\epsilon_n^2),$$

for r and M satisfying

$$r^d = \frac{2C_0}{D_2}n\epsilon_n^2, \quad M^2 = \max\{8C_0, 16C_4\}n\epsilon_n^2(\log(r/\epsilon_n))^{1+d}. \quad (4.16)$$

Denote the solution of the above equation as r_n and M_n .

By Lemma 4.4, for $M\sqrt{r/\delta} > 2\epsilon$ and $r > a_0$,

$$\begin{aligned} \log N\left(2\epsilon, M\sqrt{\frac{r}{\delta}}\tilde{\mathbb{H}}_1^r + \epsilon\tilde{\mathbb{B}}_1, \|\cdot\|_\infty\right) &\leq \log N\left(\epsilon, M\sqrt{\frac{r}{\delta}}\tilde{\mathbb{H}}_1^r, \|\cdot\|_\infty\right) \\ &\leq Kr^d \left(\log\left(\frac{M\sqrt{r/\delta}}{\epsilon}\right)\right)^{1+d}. \end{aligned}$$

By Lemma 4.7, every element of $M\tilde{\mathbb{H}}_1^a$ for $a < \delta$ is uniformly at most $\delta\sqrt{D}\tau M$ distant from a constant function for a constant in the interval $[-M, M]$. Therefore for $\epsilon > \delta\sqrt{D}\tau M$,

$$\log N\left(3\epsilon, \bigcup_{a < \delta} (M\tilde{\mathbb{H}}_1^a) + \epsilon\tilde{\mathbb{B}}_1, \|\cdot\|_\infty\right) \leq N(\epsilon, [-M, M], |\cdot|) \leq \frac{2M}{\epsilon}.$$

With $\delta = \epsilon/(2\sqrt{D}\tau M)$, combining the above displays, for $B = B_{M,r,\delta,\epsilon}$ with

$$M \geq \epsilon, \quad M^{3/2}\sqrt{2\tau r}D^{1/4} \geq 2\epsilon^{3/2}, \quad r > a_0,$$

which is satisfied when $r = r_n$ and $M = M_n$, we have

$$\log N(3\epsilon, B, \|\cdot\|_\infty) \leq Kr^d \left(\log\left(\frac{M^{3/2}\sqrt{2\tau r}D^{1/4}}{\epsilon^{3/2}}\right)\right)^{1+d} + \log \frac{2M}{\epsilon}. \quad (4.17)$$

Therefore, for $r = r_n$ and $M = M_n$,

$$\log N(3\bar{\epsilon}_n, B, \|\cdot\|_\infty) \leq n\bar{\epsilon}_n^2,$$

for $\bar{\epsilon}_n$ a large multiple of $\epsilon_n(\log n)^{\kappa_2}$ with $\kappa_2 = (1+d)/2$. \square

Proof of Theorem 2.2. Under d' , the prior concentration inequality becomes:

$$\begin{aligned} P(\|W^A - f_0\|_\infty \leq 2\epsilon) &\geq P(\|W^A - f_0\|_\infty \leq 2\epsilon, A \in [(C/\epsilon)^{1/s}, 2(C/\epsilon)^{1/s}]) \\ &\geq \int_{(C/\epsilon)^{1/s}}^{2(C/\epsilon)^{1/s}} e^{-\phi_{f_0}^s(\epsilon)} g(a) da \\ &\geq C_1 e^{-K_2(1/\epsilon)^{d\vee d'/s}(\log(1/\epsilon))^{1+d}} \left(\frac{C}{\epsilon}\right)^{p/s} \left(\frac{C}{\epsilon}\right)^{1/s}. \end{aligned} \quad (4.18)$$

The complementary probability becomes:

$$P(W^A \notin B) \leq \frac{2C_2 r^{p-d'+1} e^{-D_2 r^{d'}}}{D_2} + e^{-M^2/8}, \quad (4.19)$$

with $M^2 \geq 16C_4 r^d (\log(r/\epsilon))^{1+d}$, $r > 1$ and $\epsilon < \epsilon_1$, where $\epsilon_1 > 0$ is a fixed constant.

An upper bound for the covering entropy is unchanged and still given by (4.17).

1. $d' > d$: With ϵ_n a multiple of $n^{-s/(2s+d')} (\log n)^{\kappa_1}$ with $\kappa_1 = (1+d)/(2+d'/s)$, $\bar{\epsilon}_n < \epsilon_n$,

$$r^{d'} = \frac{2C_0}{D_2} n \epsilon_n^2, \text{ and } M^2 = \max\{8C_0, 16C_4\} n \epsilon_n^2 (\log(r/\epsilon_n))^{1+d},$$

inequalities (4.18), (4.19) and (4.17) become

$$\begin{aligned} P(\|W^A - f_0\|_\infty \leq \epsilon_n) &\geq \exp(-n \epsilon_n^2), \\ P(W^A \notin B) &\leq \exp(-C_0 n \epsilon_n^2), \\ \log N(3\bar{\epsilon}_n, B, \|\cdot\|_\infty) &\leq n \bar{\epsilon}_n^2. \end{aligned}$$

Comparing the above with (4.1), we arrive at the conclusion that under $d' > d$, the posterior contraction rate will be at least a multiple of $n^{-s/(2s+d')} (\log n)^\kappa$ with $\kappa = (1+d)/(2+d'/s)$.

2. $\frac{d^2}{2s+d} < d' < d$: With ϵ_n a multiple of $n^{-s/(2s+d)} (\log n)^{\kappa_1}$ with $\kappa_1 = (1+d)/(2+d/s)$, $\bar{\epsilon}_n$ a multiple of $n^{d/(2d')-1} \epsilon_n^{d/d'} (\log n)^{(d+1)/2} = n^{-\frac{(2s+d)d'-d^2}{2(2s+d)d'}} (\log n)^{\kappa_2}$ with $\kappa_2 = (d+d^2)/(2d'+dd'/s) + (1+d)/2$,

$$r^{d'} = \frac{2C_0}{D_2} n \epsilon_n^2, \text{ and } M^2 = \max\{8C_0, 16C_4\} n \epsilon_n^2 (\log(r/\epsilon_n))^{1+d},$$

inequalities (4.18), (4.19) and (4.17) become

$$\begin{aligned} P(\|W^A - f_0\|_\infty \leq \epsilon_n) &\geq \exp(-n \epsilon_n^2), \\ P(W^A \notin B) &\leq \exp(-C_0 n \epsilon_n^2), \\ \log N(3\bar{\epsilon}_n, B, \|\cdot\|_\infty) &\leq n \bar{\epsilon}_n^2. \end{aligned}$$

Comparing the above with (4.1), we arrive at the conclusion that under $d' < d$, the posterior contraction rate will be at least a multiple of $n^{-\frac{(2s+d)d'-d^2}{2(2s+d)d'}} (\log n)^\kappa$ with $\kappa = (d+d^2)/(2d'+dd'/s) + (1+d)/2$. To make this rate meaningful, we need $(2s+d)d' - d^2 > 0$, i.e. $d' > d^2/(2s+d)$. \square

5. Numerical Example

We provide a numerical example using the lucky cat data (Fig. 1). This data set has intrinsic dimensionality one, which is the dimension of the rotation

TABLE 1

Square root of MSPE for the lucky cat data by using three different approaches over 100 random splitting are displayed. The numbers in the parenthesis indicate the standard deviations.

	$n = 18$	$n = 36$	$n = 54$
EN	.416(.152)	.198(.042)	.149(.031)
LASSO	.431(.128)	.232(.061)	.163(.038)
GP	.332(.068)	.128(.036)	.077(.014)
2GP	.181(.051)	.124(.038)	.092(.021)
RPGP	.340(0.071)	.130(.039)	.077(.015)

angle θ . Since we know the true value of θ , we create the truth $f_0(\theta) = \cos \theta$ as a continuous function on the unit circle. The responses are simulated from $Y_i = f_0(\theta_i) + \epsilon_i$ by adding independent Gaussian noises $\epsilon_i \sim N(0, 0.1^2)$ to the true values. In this model, the total sample size $N = 72$ and the predictors $X_i \in \mathbb{R}^p$ with $D = 16,384$. To assess the impact of the sample size n on the fitting performance, we randomly divide $n = 18, 36$ and 64 samples into training set and treat the rest as testing set. Training set is used to fit a model and testing set to quantify the estimation accuracy. For each training size n , we repeat this procedure for $m = 100$ times and calculate the square root of mean squared prediction error (MSPE) on the testing set,

$$\sum_{l=1}^m \frac{1}{N-n} \sum_{i \in T_l} \|\hat{Y}_i - f_0(\theta_i)\|^2,$$

where T_l is the l th testing set and \hat{Y}_i is an estimation of $E[Y|X_i] = f_0(\theta_i)$. We apply three GP based algorithms on this data set: 1. vanilla GP specified by (2.3); 2. Two stage GP (2GP) where the D -dimensional predictors were projected into \mathbb{R}^2 by using Laplacian eigenmap [3] in the first stage and then a GP with projected features as predictors was fitted in the second stage; 3. Random projection GP (RPGP) where the new predictors were produced by projecting the original predictors into \mathbb{R}^{1000} with a random projection matrix $\Psi^P = (\Psi_{lj}) \in \mathbb{R}^{1000 \times 16384}$ with $\Psi_{lj} \sim \text{i.i.d. } N(0, 1)$. To assess the prediction performance, we also compare our GP prior based models (2.3) with lasso [28] and elastic net (EN) [34] under the same settings. We choose these two competing models because they are among the most widely used methods in high dimensional regression settings and perform especially good when the true model is sparse. In the GP models, we set $d = 1$ since the sample size for this dataset is too small for most dimension estimation algorithms to reliably estimate d . In addition, for each simulation, we run 10,000 iterations with the first 5,000 as burn-in.

The results are shown in Table. 1. As we can see, under each training size n , GP performs the best. Moreover, as n increases, the prediction error of GP decays much faster than EN and Lasso: when $n = 18$, the square root of MSPEs by using EN and lasso are about 125% of that by using GP; however as n increases to 54, this ratio becomes about 200%. Moreover, the standard deviation of square root of MSPEs by using GP are also significantly lower than those

TABLE 2

Square root of MSPE for the lucky cat data with noised predictors. results over 100 random splitting are displayed. The numbers in the parenthesis indicate the standard deviations. The numbers after RPGP indicates the projected dimension \tilde{d} .

σ_X	0	10	20	40	80
GP	.077(.014)	.095(.015)	.116(.017)	.180(.020)	.276(.23)
RPGP(10)	.275(.065)	.291(.069)	.335(.075)	.452(.085)	.606(.102)
RPGP(100)	.106(.023)	.116(.026)	.143(.033)	.225(.043)	.360(.065)
RPGP(1000)	.077(.015)	.088(.017)	.102(.018)	.178(.021)	.289(.033)

by using lasso and EN. Among GP based methods, RPGP has slightly worse performance than GP under small training size, but as n grows to 54, they have comparable MSPEs. It is not surprising that 2GP has better performance than GP when n is small since the dimensionality reduction map Ψ is constructed using the whole dataset (the Laplacian eigenmap code we use cannot do interpolations). Therefore when the training size n become closer to the total data size 72, GP becomes better. In addition, GP is computationally faster than 2GP due to the manifold learning algorithm in the first stage of 2GP.

To compare the performances between GP and RPGP in the case when there are noises in the predictors, we add $N(0, \sigma_X I_D)$ noises into each predictor vector X_i with noise levels $\sigma_X = 0, 10, 20, 40$ and 80, where the range of predictors is $0 \sim 255$. We also change the projected dimension \tilde{d} from 10 to 1,000. The training size n is fixed at 54. Table. 2 displays the results.

As we can see, for small $\tilde{d} = 10$ or 100, applying GP on the original predictors appears to be better than RPGP on the projected predictors under any settings. As \tilde{d} grows to 1,000, GP and RPGP have similar performances in the noise free setting. However, as noises are added to the predictors, RPGP with $\tilde{d} = 1,000$ outperforms GP. However, as the noise increases to the order comparable to the signals, GP becomes close to and finally outperforms RPGP. In addition, the standard deviation of RPGP also grows rapidly as noise increases. This suggests that GP might be more stable than RPGP under small signal-to-noise ratio scenarios.

6. Discussion

In this work, we considered a nonparametric Bayesian prior for high dimensional regression when the predictors are assumed to be lying on a low dimensional intrinsic manifold. The proposed prior can be considered as an extension of a Gaussian process prior on Euclidean space to a general submanifold. We show that this GP prior can attain near optimal posterior convergence rate that can adapt to both the smoothness of the true function ($s \leq 2$) and the underlying intrinsic manifold \mathcal{M} . Our theorem validates the surprising phenomenon suggested by Bickel in his 2004 Rietz lecture [5] under the GP prior scenario:

“... the procedures used with the expectation that the ostensible dimension D is correct will, with appropriate adaptation not involving manifold estimation, achieve the optimal rate for manifold dimension d .”

Moreover, we also provide theoretical guarantees for two stage GP with dimensionality reduction. We suggest the use of random projection GP as a special two stage GP when noises exist in the predictors.

One possibility of our future work is to investigate whether the smoothness requirement $s \leq 2$ could be relaxed. This extension will be dependent on whether Lemma 4.2 could be improved to $s \geq 2$. Currently we construct the approximation function $I_a(f)$ in RKHS through convolving f with the covariance function. It is not clear whether this is the best way to approximate the function f by elements in the RKHS.

A second possibility is to build a coherent model not only estimating the regression function $E[Y|X]$, but simultaneously learning the dimensionality d of the intrinsic manifold \mathcal{M} . Our current GP prior (2.3) completely ignores the information contained in the marginal distribution P_X of the predictor X . As an alternative, we can only model part of P_X and therefore utilize some of P_X 's information, such as the support or dimensionality of \mathcal{M} .

References

- [1] ARONSZAJN, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society* **68** 337-404.
- [2] BARANIUK, R. G. and WAKIN, M. B. (2009). Random projections of smooth Manifolds. *Found. Comput. Math.* **9** 51-77.
- [3] BELKIN, M. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* **15** 1373-1396.
- [4] BHATTACHARYA, A., PATI, D. and DUNSON, B. D. Adaptive dimension reduction with a Gaussian process prior. *arXiv: 1111.1044*.
- [5] BICKEL, J. P. and LI, B. (2007). Local polynomial regression on unknown manifolds. *Complex datasets and inverse problem: tomography, networks and beyond, IMS Lecture Notes-Monograph Series*, **54** 177-186.
- [6] CAMASTRA, F. and VINVIARELLI, A. (2002). Estimating the intrinsic dimension of data with a fractal-based method. *IEEE P.A.M.I.* **24** 1404-1407.
- [7] CARTER, K. M., RAICH, R. and HERO, A. O. (2010). On local intrinsic dimension estimation and its applications. *Trans. Sig. Proc.* **58** 650-663.
- [8] CHEN, M., SILVA, J., PAISLEY, J., WANG, C., DUNSON, D. B. and CARIN, L. (2010). Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. *IEEE Trans. Signal Process* **58** 6140-6155.
- [9] DO CARMO, M. (1992). *Riemannian geometry*. Birkhauser, Boston.
- [10] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500-531.
- [11] GINE, E. and KOLTCHINSKII, V. (2005). Empirical graph Laplacian approximation of Laplace Beltrami operators: Large sample results. *High Dimensional Probability IV* **51** 238-259.

- [12] HIAI, F. and PETZ, D. (2009). Riemannian metrics on positive definite matrices related to means. *Linear Algebra and its Applications* **430** 3105-3130.
- [13] KENT, J. T. (1992). New directions in shape analysis In *The Art of Statistical Science. A Tribute to G. S. Watson* 115-127. Wiley, New York.
- [14] KUELBS, W. V. J. ABD LI and LINDE, W. (1994). The Gaussian measure of shifted balls. *Probab. Theory Related Fields* **98** 143-162.
- [15] KUNDU, S. and DUNSON, D. B. (2011). Latent factor models for density estimation. *arXiv:1108.2720v2*.
- [16] LAWRENCE, N. D. (2003). Gaussian process latent variable models for visualisation of high dimensional data. *Neural Information Processing Systems*.
- [17] LEVINA, E. and BICKEL, P. (2004). Maximum likelihood estimation of intrinsic dimension In *Advances in Neural Information Processing Systems* **17**. The MIT Press, Cambridge, MA, USA.
- [18] LITTLE, A. V., LEE, J., JUNG, Y. M. and MAGGIONI, M. (2009). Estimation of intrinsic dimensionality of samples from noisy low-dimensional manifolds in high dimensions with multiscale SVD In *2009 IEEE/SP 15th Workshop on Statistical Signal Processing* 85-88.
- [19] MOAKHER, M. and ZÉRAÏ, M. (2011). The Riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data. *Journal of Mathematical Imaging and Vision* **40** 171-187.
- [20] NASH, J. (1956). The imbedding problem for Riemannian manifolds. *Annals of Mathematics* **63** 20-63.
- [21] NENE, S. A., NAYAR, S. K. and MURASE, H. (1996). Columbia object image library (COIL-100) Technical Report, Columbia University.
- [22] PAGE, G., BHATTACHARYA, A. and DUNSON, D. B. (2013). Classification via Bayesian nonparametric learning of affine subspaces. *J. Amer. Statist. Assoc.* **108** 187-201.
- [23] REICH, B. J., BONDELL, H. D. and LI, L. X. (2011). Sufficient dimension reduction via Bayesian mixture modeling. *Biometrics* **67** 886-895.
- [24] ROWEIS, S. T. and SAUL, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** 2323-2326.
- [25] SAVITSKY, T., VANNUCCI, M. and SHA, N. (2011). Variable selection for nonparametric Gaussian process priors: models and computational strategies. *Statistical Science* **26** 130-149.
- [26] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040-1053.
- [27] TENENBAUM, J. B., SILVA, V. and LANGFORD, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319-2323.
- [28] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73** 273-282.
- [29] TOKDAR, S. T., ZHU, Y. M. and GHOSH, J. K. (2010). Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian*

- Anal.* **5** 319-344.
- [30] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. *IMS Collections* **3** 200-222.
 - [31] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *Ann. Statist.* **37** 2655-2675.
 - [32] VAN DER VAART, A. W. and WELLNER, J. A. (2000). *Weak convergence and empirical processes: with applications to statistics, 2nd ed.* Springer.
 - [33] YE, G. and ZHOU, D. (2008). Learning and approximation by Gaussians on Riemannian manifolds. *Adv. Comput. Math.* **29** 291-310.
 - [34] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 301-320.